Likestillings- og diskrimineringsombudet

VEILEDER

Innebygd diskrimineringsvern

En veileder for å avdekke og forebygge diskriminering i utvikling og bruk av kunstig intelligens

Innebygd diskrimineringsvern

En veileder for å avdekke og forebygge diskriminering i utvikling og bruk av kunstig intelligens

Likestillings- og diskrimineringsombudet

LDO 2023

Forfatter: Kathinka Theodore Aakenes Vik

Forsidefoto: Tzido

Utforming: Munch design

978-82-8320-027-0 (trykt utgave)

978-82-8320-028-7 (elektronisk utgave)

Innhold

Om denne veilederen

Målgruppe Ansvarlige for utvikling, innkjøp og bruk av maskinlæringssystemer der bruken tilsier at systemene kan få betydning for personers rettigheter og plikter.

Fundament Veilederen tar utgangspunkt i likestillings- og diskrimineringsloven og menneskerettslige prinsipper om likeverd, autonomi og rettferdighet.

Struktur Veilederen er tredelt:

DEL 1 omhandler formålet med denne veilederen og diskrimineringslovverkets relevans ved bruk av maskinlæringssystemer.
DEL 2 forklarer hva «diskriminering» rettslig sett er, og demonstrerer konkret hvordan maskinlæringssystemer kan virke diskriminerende.
DEL 3 skisserer relevante diskrimineringsrisikoer med tilhørende spørsmål som bør drøftes for å avdekke risiko for diskriminering. Spørsmålene er strukturert etter fem faser for utvikling og bruk av teknologien. Dette gir et utgangspunkt for å kartlegge systemet i et diskrimineringsperspektiv, slik at de ansvarlige kan iverksette tiltak for å redusere risikoen for diskriminering.

Bidrag Et diskusjons- og beslutningsverktøy egnet for selvrapportering som legger til rette for en tverrfaglig dialog mellom involverte aktører, med mål om å fremme likestilling og hindre diskriminering.

DEL 1

Bakgrunn for veilederen

Det siste tiåret har det gjentatte ganger blitt avdekket at algoritmer og maskinlæringsmodeller har ført til diskriminering, ofte fordi teknologien ikke er grundig nok vurdert. To eksempler på dette er ansiktsgjenkjenning som ikke fungerer tilstrekkelig på personer med mørkere hudfarge,[1] og algoritmer brukt til å avdekke svindel som feilaktig utpeker personer med innvandrerbakgrunn.[2]

Eksemplene viser at det ikke er tilstrekkelig å ha diskrimineringsrisiko i bakhodet ved utvikling og bruk av denne teknologien. Diskriminering må forebygges og måles systematisk.[3] Tiltak for å forebygge diskriminering og fremme likestilling må bygges inn i alle utviklingsfasene i et maskinlæringssystem (ML-system), fra planlegging til bruk av teknologien. Dette kaller vi *innebygget diskrimineringsvern*.

Formålet med veilederen

Formålet med veilederen er at de ansvarlige for utvikling, anskaffelse og bruk av ML-systemer skal gjøres kjent med diskrimineringsregelverket, og at de skal kunne forebygge diskriminering ved å vurdere diskrimineringsrisikoen teknologien kan medføre.

Behovet for denne kunnskapen er prekært, ettersom forskning tilsier at det i offentlig sektor er lav bevissthet og begrenset kompetanse om diskrimineringsregelverket i forbindelse med utvikling og bruk av denne teknologien.[4]

Veilederen retter seg primært mot ML-systemer som benyttes internt i virksomheter, ofte som beslutningsstøtte, fremfor digitale tjenester som er ment å bli brukt direkte av tjenestemottakeren, som f.eks. digitale læremidler. Imidlertid kan noen av utfordringene og spørsmålene som skisseres i DEL 3, også være relevante for at digitale tjenester skal utformes på en inkluderende måte.

1 Blant annet: https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/ og https://www.media.mit.edu/articles/if-you-re-a-darker-skinned-woman-this-is-how-often-facial-recognition-software-decides-you-re-a-man/.

2 Blant annet SyRI-skandalen i Nederland, se https://algorithmwatch.org/en/syri-netherlands-algorithm/.

3 Se blant annet Europaparlamentets forslag til EUs AI Act artikkel 29 a, som stiller krav om «fundamental rights impact assessments» (fritt oversatt til «menneskerettslige konsekvensvurderinger») ved bruk av høyrisiko kunstig intelligens, tilgjengelig: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html.

4 «Bruk av kunstig intelligens i offentlig sektor og risiko for diskriminering – Kunnskapsgrunnlag for arbeidet med å forebygge diskriminerende effekter ved bruk av kunstig intelligens i offentlig virksomhet» av Hilde G. Corneliussen, Aisha Iqbal, Gilda Seddighi og Rudolf Andersen, av desember 2022, tilgjengelig: https://www.vestforsk.no/sites/default/files/2023-03/VFrapport7_2022_KI_i_offentlig_sektor.pdf.

Diskrimineringsrettens relevans ved bruk av kunstig intelligens

Kunstig intelligens som påvirker beslutninger om personers rettsstilling

Kunstig intelligente systemer, herunder ML-systemer, tas i bruk på stadig flere samfunnsområder. Blant annet blir prediktive ML-modeller tatt i bruk i offentlig og privat sektor på en måte som er egnet til å påvirke beslutninger om personers rettigheter og plikter.

Når denne teknologien tas i bruk, er ofte formålet å fatte mer presise avgjørelser på en mer effektiv måte. Dette kan avgjøre om en borger skal tildeles ulike velferdsgoder eller pålegges plikter, og avgjørelsen kan skje på bakgrunn av *personlige kjennetegn som personen deler med andre* med lignende kjennetegn.

Dette kan stå i konflikt med forbudet mot å diskriminere. Likestillings- og diskrimineringsloven forbyr *ulovlig forskjellsbehandling* av *bestemte grupper*.[5] Loven har som formål å sikre et samfunn der alle kan delta på like vilkår og bli vurdert ut ifra sine *individuelle evner og behov*.

Typiske diskrimineringsutfordringer ved bruk av kunstig intelligens

Maskinlæringsteknologi har en iboende utfordring ved at systemenes presisjon avhenger av tilgang til et omfattende datasett. Data om *minoritetsgrupper* er ofte mindre omfangsrike, nettopp fordi gruppen utgjør en minoritet. Derfor er det et typisk problem at teknologien fungerer mindre presist for minoriteter – som igjen er grupper diskrimineringsretten verner om.

Når ML-systemer brukes til å forskjellsbehandle, er ofte det diskrimineringsrelevante spørsmålet om systemet:

- *Enten:* vurderer enkelte grupper strengere enn andre, f.eks. Amazons rekrutteringsalgoritme omtalt under.[6]
- *Eller:* fungerer dårligere eller mindre presist for enkelte grupper, f.eks. ansiktsgjenkjenning som ofte fungerer dårligere for personer med mørk hudfarge,[7] eller AHUS' hjertesviktalgoritme omtalt under.[8]

Teknologinøytralt lovverk

Likestillings- og diskrimineringsloven er teknologinøytral. Det betyr at forbudet mot diskriminering gjelder uavhengig av om den diskriminerende praksisen eller avgjørelsen er tatt av en person eller følger av et kunstig intelligent system. Slik sett regulerer forbudet mot diskriminering kun sluttresultatet, altså den forskjellsbehandlingen ML-systemet bidrar til.[9] Samtidig er det en rekke faktorer ved både utvikling og bruk av denne teknologien som *øker risikoen for at systemet bidrar til et diskriminerende utfall*. Disse risikofaktorene omtales under de fem fasene i DEL 3.

5 Diskrimineringsgrunnlagene er kjønn, graviditet, permisjon ved fødsel eller adopsjon, omsorgsoppgaver, etnisitet, religion, livssyn, funksjonsnedsettelse, seksuell orientering, kjønnsidentitet, kjønnsuttrykk, alder eller kombinasjoner av disse grunnlagene, jf. likestillings- og diskrimineringsloven (forkortet ldl.) § 6.

6 DEL 3, fase 4 om testing av systemet.

7 Se blant annet: https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf og https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/.

8 Se DEL 3, fase 1 og 5.

9 For offentlig sektor strekker forpliktelsen seg lenger enn selve diskrimineringsforbudet, se DEL 2 om Likestillingsplikten i offentlig sektor.

DEL 2

Hva er diskriminering?

Norsk lov forbyr diskriminering på alle samfunnsområder.[10] Forbudet bygger på likhets- og ikke-diskrimineringsprinsippet som er forankret i Grunnloven § 98. Dette prinsippet står også sentralt i Den europeiske menneskerettskonvensjon (EMK) og flere andre menneskerettskonvensjoner.[11]

Diskriminering etter likestillings- og diskrimineringsloven (ldl.) kan defineres som:

Ulovlig forskjellsbehandling knyttet til et eller flere diskrimineringsgrunnlag.[12]

Noen eksempler på ulovlig forskjellsbehandling kan være:

- Et forsikringsselskap tilbyr konsekvent høyere forsikringspriser til kunder med innvandrerbakgrunn
- Et boligutleieselskap sorterer ut søknader fra interessenter fordi de har besøkt nettsider som retter seg mot homofile
- Politiet arresterer feil person fordi ansiktsgjenkjenningsteknologien ikke fungerer tilstrekkelig på personer med mørk hudfarge
- En arbeidsgiver benytter seg ukritisk av en rekrutteringsalgoritme som kun anbefaler norske menn mellom 18 og 30 år og undervurderer andre grupper
- En blind person innvilges uføretrygd automatisk og mister retten til jobbsøkerkurs.

Det å ikke få tildelt en utleiebolig, ikke få tilbud om en jobb eller å bli arrestert er avgjørelser som innebærer at noen stilles dårligere enn andre. Hvorvidt det er et menneske, en arbeidsgiver, et selskap eller et maskinlæringssystem som står bak disse avgjørelsene, er i utgangspunktet uten betydning for om noe kan være diskriminering. Det avgjørende er virkningen forskjellsbehandlingen har for personen som rammes.[13]

Forbudet mot å diskriminere betyr ikke nødvendigvis at alle skal behandles på akkurat samme måte. Tvert imot legger loven føringer for tilrettelegging for enkelte med slike behov for å nyttiggjøre seg av en tjeneste. Eksempler er individuelt tilrettelagt undervisningsopplegg som kan være nødvendig for å sikre utbytte av skolegang, og digitale flater som må tilpasses synshemmede for å være tilgjengelige for alle.[14]

10 Ldl. §§ 2 og 6.

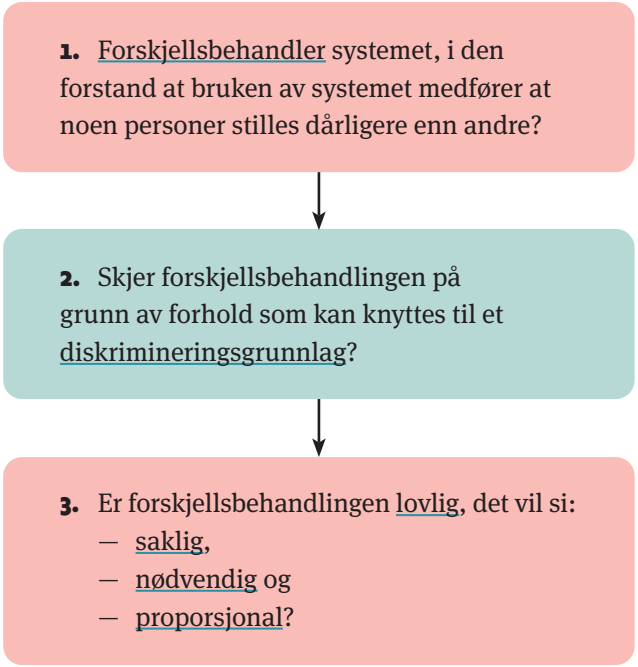
11 EMK art. 14, Den internasjonale konvensjonen om sivile og politiske rettigheter (SP) av 16. desember 1966, art. 2 første ledd og art. 26, Den internasjonale konvensjonen om økonomiske, sosiale og kulturelle rettigheter (ØSK) av 16. desember 1966, art. 2 andre ledd, Den europeiske menneskerettskonvensjonen med protokoller (EMK) av 4. november 1950, art. 14 og FNs konvensjon av 20. november 1989 om barnets rettigheter med protokoller (Barnekonvensjonen), art. 2 første ledd.

12 Ldl. §§ 6, 7, 8, 9.

13 Det er kun fysiske personer, og ikke juridiske personer, som har et diskrimineringsvern etter ldl., jf. lovens forarbeider, Prop. 81 L (2016–2017) avsnitt 10.9.

14 Jf. bestemmelsene om universell utforming og individuell tilrettelegging i ldl. §§ 12, 17, 18, 20, 21, 22, 23, samt generell bestemmelse om indirekte forskjellsbehandling i ldl. § 8.

Det er tre spørsmål som må vurderes for å avgjøre om noe utgjør diskriminering, slik begrepet er definert over:



1. spørsmål:
Forskjellsbehandling

Forskjellsbehandling betyr at en person blir behandlet dårligere enn andre i en tilsvarende situasjon.[15] Det innebærer for det første at forskjellsbehandlingen må føre til skade eller ulempe for den som forskjellsbehandles, for eksempel at forskjellsbehandlingen fører til tap av fordeler, økonomisk tap eller færre muligheter sammenlignet med andre som er i en tilsvarende situasjon.

Ved spørsmål om et ML-system forskjellsbehandler, er det avgjørende å vurdere hele tjenesten som systemet inngår i som komponent.

Forskjellsbehandlingen kan skje *direkte* eller *indirekte*.[16] At forskjellsbehandlingen rammer personen *direkte*, vil si at personen behandles dårligere *på grunn av* vedkommendes kjønn, etnisitet, funksjonsevne eller livssyn mm. Forskjellsbehandling som rammer *indirekte*, defineres som en tilsynelatende nøytral praksis eller avgjørelse som er egnet til å sette noen grupper i en dårligere situasjon enn andre.

Rekrutteringsalgoritmen som eksempel på direkte og indirekte forskjellsbehandling

Et illustrerende eksempel er en rekrutteringsalgoritme som gir en lavere score til kvinnelige arbeidssøkere i en rekrutteringsprosess på grunn av søkerens kjønn. Dette er *direkte forskjellsbehandling*.[17]

Dersom algoritmen gir lavere score til søkere med arbeidserfaring fra kvinnedominerte bransjer, vil ikke søkeren behandles dårligere direkte *på grunn av* søkerens kjønn, men algoritmens vekting av erfaring er likevel *egnet til å ramme kvinner* på en måte som gjør at kvinnelige søkere stilles dårligere enn menn. Dette er et eksempel på *indirekte forskjellsbehandling*.[18]

2. spørsmål:
Diskrimineringsgrunnlag

Det andre spørsmålet for at noe skal regnes som diskriminering, er om det finnes en sammenheng mellom forskjellsbehandlingen og et *diskrimineringsgrunnlag*. Kort fortalt gjenspeiler diskrimineringsgrunnlagene bestemte forhold ved en person.

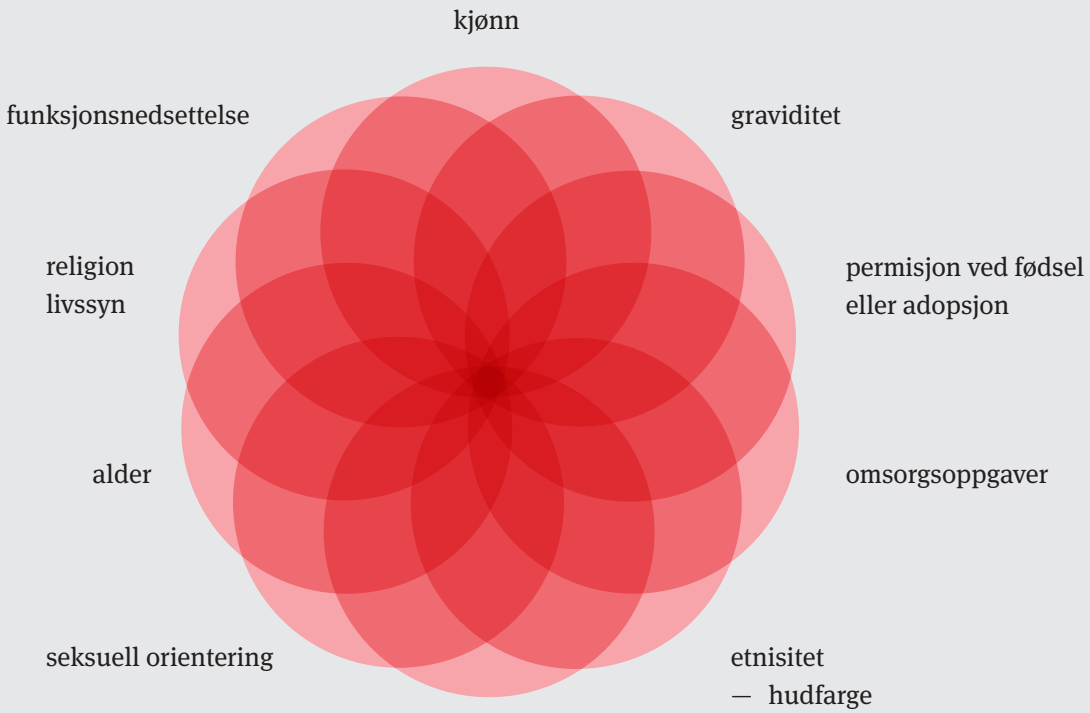
15 Ldl. §§ 7 og 8.

16 Prop. 81 (2016–2017) avsnitt 12.2.4.1, 101–102.

17 Ldl. § 7.

18 Ldl. § 8.

De aktuelle diskrimineringsgrunnlagene som er definert i loven, er:



At det er akkurat disse grunnlagene som er nedfelt i loven, er ikke tilfeldig. De er ment å fange grupper som historisk har vært og fremdeles ofte blir dårligere behandlet enn andre. Disse grunnlagene for diskriminering har ofte å gjøre med forhold eller egenskaper som den diskriminerte ikke kan endre.

Bruk av tilsynelatende nøytrale opplysninger som sammenfaller med diskrimineringsgrunnlag

Som nevnt under omtalen av det første spørsmålet i diskrimineringsvurderingen, kan forskjellsbehandlingen skje både *direkte* og *indirekte*. Dersom rekrutteringsalgoritmen som nevnt over, gir lavere score til personer avhengig av *bosted*, vil dette ikke fanges opp av diskrimineringsloven, fordi bosted ikke er et såkalt *vernet grunnlag*.

Imidlertid kan det likevel tenkes at *bosted* delvis overlapper med *etnisitet* dersom det er flere personer med innvandrerbakgrunn bosatt i bestemte bydeler, og dermed kan vektleggingen av en tilsynelatende *nøytral opplysning* som bosted bidra til indirekte diskriminering på grunn av etnisitet. Denne problematikken er særlig relevant ved ML-systemer og er ytterligere omtalt under DEL 3 fase 4 om testing av systemet.

Sammensatt diskriminering

Et fenomen som er særlig relevant ved bruken av ML-systemer, er det som omtales som *sammensatt diskriminering*. Med dette menes diskriminering på flere grunnlag samtidig, eller at kombinasjonen av flere diskrimineringsgrunnlag utgjør grunnlaget for diskriminering.

Dette betyr i praksis at forskjellsbehandlingen ikke nødvendigvis er synlig på et avgrenset gruppenivå. Ett eksempel er dersom *kvinner* og *menn* tilsynelatende behandles likt på gruppenivå, men der *menn* med *asiatisk* opphav i en gitt *alder* likevel blir dårligere behandlet på grunn av de nevnte kjennetegnene.

Dette kan være vanskeligere å oppdage fordi det kan tenkes mange ulike kombinasjoner av undergrupper, og det kreves målrettet testing av systemet for en snevrere gruppe enn det diskrimineringsgrunnlagene skulle tilsi.

19 Arbeidsmiljøloven § 13-1

3. spørsmål:
«Ulovlig»

Det tredje og siste spørsmålet for å avgjøre om noe er *diskriminering* i lovens forstand, er om forskjellsbehandlingen er *ulovlig*. Det kan være gode grunner til at noen personer stilles dårligere enn andre, og slik forskjellsbehandling kan være tillat dersom den:

— har et *saklig formål*
— er *nødvendig* for å oppnå dette formålet
— og *proporsjonal*, altså at det er et rimelig forhold mellom det man ønsker å oppnå og hvor inngripende forskjellsbehandlingen er for den som rammes.

Dersom samtlige tre vilkår for *lovlig* forskjellsbehandling er oppfylt, vil ikke forskjellsbehandlingen regnes som diskriminering.

Rekrutteringsalgoritmen som eksempel på saklig og usaklig forskjellsbehandling

Dersom en rekrutteringsalgoritme sorterer ut søkere som er under 18 år til en stilling som krever universitetsutdanning, kan det være gode grunner for at denne forskjellsbehandlingen på grunn av alder likevel er lovlig.

Men dersom den sorterer ut søkere som har en funksjonsnedsettelse, eller søkere som oppgir et annet morsmål enn norsk, vil dette trolig ikke ha et saklig formål, være nødvendig, og ei heller være proporsjonalt. I slike tilfeller må man gjøre en individuell vurdering av søkeren for å avgjøre om det kan være saklig å forskjellsbehandle. Hvis algoritmen automatisk utelukker søkere, får man ikke gjort en slik individuell vurdering, og faren for diskriminering er betydelig.

Oppsummering

Dersom ML-systemet:

1. forskjellsbehandler, ved at bruken av systemet medfører at noen personer stilles dårligere enn andre
2. forskjellsbehander på grunn av forhold som kan knyttes til et diskrimineringsgrunnlag
3. forskjellsbehandlingen ikke er saklig, nødvendig eller proporsjonal

vil systemet kunne legge grunnlag for diskriminerende avgjørelser.

De færreste ML-systemer er per nå autonome, og tar sjelden selvstendige avgjørelser. Ofte er slike systemer underlagt menneskelig kontroll. Dersom systemet brukes som ledd i en beslutningsprosess, må eieren av systemet kartlegge om systemet diskriminerer enkelte persongrupper, og kompensere for eventuelle diskriminerende tendenser i systemet. Mer om dette i forbindelse med DEL 3, fase 5 om implementering.

Særlig om likestillingsplikten i offentlig sektor

I tillegg til det generelle forbudet mot å diskriminere er offentlige myndigheter ytterligere forpliktet til å jobbe aktivt, målrettet og planmessig for å fremme likestilling og hindre diskriminering i all sin myndighetsutøvelse og all tjenesteyting.[20]

Plikten innebærer at det må stilles særlige krav til maskinlæringssystemer som utvikles for eller tas i bruk av offentlig sektor, der systemene kan få betydning for personers rettsstilling. Av den grunn har offentlig sektor et særlig ansvar for å sikre at ML-systemer som utvikles eller tas i bruk, *fremmer likestilling og forebygger diskriminering* utover at sluttresultatet ved bruk av teknologien ikke skal diskriminere. Dette arbeidet skal redegjøres for i årlige rapporter, og dokumentene skal være offentlig tilgjengelige.

20 Ldl. § 24 om offentlige myndigheters aktivitets- og redegjørelsesplikt.

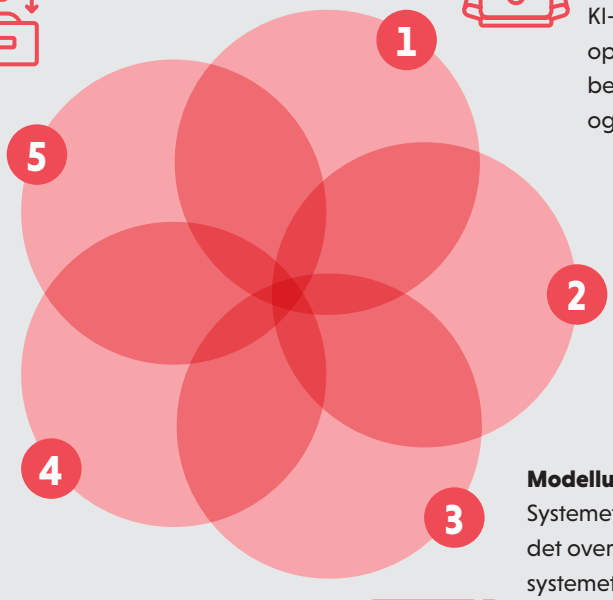
DEL 3

Typiske diskrimineringsutfordringer og relevante spørsmål til 5 faser ved utvikling og bruk av ML-systemer

Veilederen er inndelt etter fem ulike faser ved planlegging, utvikling og implementering inkludert oppfølging av et ML-system. Under hver fase skisseres diskrimineringsrettslig relevante problemstillinger med tilhørende spørsmål som kan bidra til å avdekke diskrimineringsrisiko. Svarene på spørsmålene bør dokumenteres for å synliggjøre vurderingene som er gjort overfor de berørte og eventuelle tilsynsmyndigheter. For systemer som tas i bruk i offentlig sektor, kan dokumentasjonen brukes i forbindelse med likestillingsredegjørelsen som leveres i årsrapporten (jf. ldl. § 24).

Som illustrasjonen under viser, kan de fem fasene delvis gli over i hverandre. Det kan være nyttig å ha et blikk på de øvrige fasene gjennom kartleggingen av de enkelte fasene.

Det kan være utfordrende å avdekke alle tenkelige diskrimineringskonsekvenser på egen hånd. Diskriminering fra maskinlæringssystemer er sjeldent en bevisst handling, men skjer ofte utilsiktet, og kan være krevende å avdekke i utviklingsløpet. For å være best mulig rustet anbefaler vi en tverrfaglig tilnærming som trekker på ulike fagdisipliner, herunder samfunnsvitenskapelig og juridisk kompetanse, i tillegg til forankring hos beslutningstakere. Dersom virksomheten mangler kompetanse internt, bør eksterne involveres.



Planlegging
Definer formål og tiltenkt bruk av KI-systemet. Konteksten systemet skal operere i, er avgjørende for hvilken betydning det kan ha på individ- og samfunnsnivå.

Treningsdata
Datagrunnlaget modellen er trent på, må være representativt ut fra modellens formål, og historiske skjevheter må kartlegges og korrigeres.

Modellutvikling
Systemets beregning må sammenfalle med det overordnede formålet. De variablene som systemets beregninger baserer seg på, må være relevante. Påvis saklig formål, nødvendighet og forholdsmessighet for variabler som kan knyttes til et diskrimineringsgrunnlag.

Testing
Kartlegg om systemet fungerer like presist for alle grupper. Selv om personopplysninger fjernes, kan systemet finne mønstre i dataene slik at opplysninger som sammenfaller med diskrimineringsgrunnlagene vektlegges ulovlig.

Implementering og tilsyn
Undersøk om skjevheter i systemet kan utjevnes ved kompenserende tiltak. Menneskelig kontroll med systemet må sikres, og kunnskap om systemet gjøres kjent for de berørte, slik at de kan ivareta sine rettigheter.

1. Planlegging

Den første fasen handler om å sette søkelys på problemet som systemet skal løse, definere hva som er målet, suksesskriterier og hvilke sosiale og politiske konsekvenser bruken kan ha i et likestillings- og diskrimineringsperspektiv.

Relevante diskrimineringsutfordringer

ML-modellens tiltenkte bruk kan ha avgjørende betydning for diskrimineringsrisiko. Dersom ML-modellen utvikles for å beregne stjernehimmelen og planetens lokalisering, er denne veilederen sannsynligvis irrelevant. Hvis ML-modellen utvikles for å forutsi hvilke borgere som skal få innvilget velferdsgoder, er derimot spørsmålet om diskriminering aktuelt.

Videre, innenfor de modellene som kan få betydning for personers rettigheter og plikter, er det en forskjell på ML-systemer som brukes til:

I. **Kontrollformål:** Systemet har til hensikt å kontrollere, og kan brukes til å sanksjonere personer. Systemet kan diskriminere ved å kontrollere enkelte deler av befolkningen i uforholdsmessig omfang eller mer inngående enn andre.[21]
II. **Tildelingsformål:** Systemet har til hensikt å gi borgerne bedre og riktige tjenester. Slike systemer kan diskriminere dersom de fungerer mindre presist for enkelte grupper. Dette kan for eksempel skje innenfor helsesektoren ved at enkelte persongrupper ikke mottar *likeverdige offentlige tjenester* som de har krav på.

Aktuelle problemstillinger som bør drøftes

Formuler eksplisitte svar på spørsmålene under. Det hjelper virksomheten til å få opp avgjørende informasjon om risiko for diskriminering. Noen spørsmål kan besvares kort, andre krever kartlegging og samfunnsvitenskapelig, juridisk og teknisk analyse. En plan for oppfølging kan også være aktuelt.

Modellens formål:

a. Hva er systemets tiltenkte bruk?
b. Hvilken kontekst skal det operere i?
c. I hvilken grad er systemet autonomt?
d. Hvilke persongrupper differensieres og hvorfor?
e. Har representanter fra disse persongruppene blitt involvert og hørt ved planlegging og utforming av systemet?

Modellens effekt:

f. Hvilken betydning kan systemet få for enkelte personer? Brukes det (som ledd) for å avgjøre enkeltpersoners rettsstilling?
    i. **Kontrollformål:** Kontroll av personer? Som å forutsi mulig fremtidig adferd?
        1. Hvem rammes av modellen?
        2. Vil noen persongrupper være ekstra utsatt for diskriminering ved kontroll?
    ii. **Tildelingsformål:** Forbedre (tilgangen til) tjenester for personer?
        1. Hvilke persongrupper vil bli berørt av modellen?
g. Hvilken betydning kan systemet få på samfunnsnivå?
h. Kan systemet forbedre tidligere praksis med tanke på forekomst av diskriminering eller andre typer feil?

Suksesskriterier:

i. Hvordan skal suksess måles med hensyn til effektivitet eller økt presisjon?
j. Hva betyr suksesskriteriene for ulike persongrupper?

21 Blant annet skandalen i Nederland hvor ML-systemet «SyRI» ble brukt for å identifisere trygdesvindel, se https://academic.oup.com/hrlr/article/22/2/ngac010/6568079?login=false.

Lånekassens bokontroll som eksempel på kontrollformål

Lånekassen har tidligere benyttet en maskinlæringsmodell for utvelgelsen av stipendmottakere til bokontroll. I samarbeid med Riksrevisjonen ble det kommentert at Lånekassen burde vurdert spørsmål om likebehandling i utvikling av ML-modellen, og det ble videre oppdaget at menn i større grad enn kvinner ble plukket ut som «sannsynlige misligholdere» enn det var grunnlag for i datamaterialet.[22]

AHUS' hjertesviktalgoritme som eksempel på tildelingsformål

Et eksempel på bruk av kunstig intelligens til tildelingsformål er AHUS' hjertesviktalgoritme. AHUS deltok i Datatilsynets regulatoriske sandkasse for etisk kunstig intelligens i 2022, hvor de også mottok vår veiledning for å håndtere utfordringen med algoritmeskjevhet.[23] Algoritmen hadde som formål å sette sykehuset i bedre stand til å stille riktig diagnose på en mer effektiv måte, og dermed kunne gi pasientene bedre helsehjelp enn i dag. Eksempelet vil bli omtalt mer inngående under andre relevante faser.

2. Treningsdata

Denne fasen handler om å sikre forsvarlig innsamling, bearbeiding og bruk av data. Å trene opp en maskinlæringsmodell krever store mengder data, og det er på bakgrunn av disse dataene at maskinen lærer å gjenkjenne mønstre. Dataene blir avgjørende for hvilke sammenhenger systemet oppdager, og hvilke prediksjoner systemet gir.

Relevante diskrimineringsutfordringer

Datagrunnlaget som brukes til å utvikle ML-systemet, kan inneholde ulike former for skjevheter.[24] Skjevhetene kan utgjøre en sentral kilde til diskriminering i ML-systemer. En vanlig utfordring er at noen grupper er underrepresentert i treningsdataene, noe som fører til at modellen gir dårligere prediksjoner for disse gruppene. Denne formen for skjevhet omtales som *representasjonsskjevhet*.[25]

Treningsdataene som brukes, må være representative opp mot modellens formål. I den forbindelse er det avgjørende å kartlegge hvilke grupper som kan tenktes å ha et avvikende mønster fra flertallet i den konkrete beregningen systemet gjør.

Problemet med et mangelfullt datasett illustreres av generativ AI og bildegeneratorene som ble gjort tilgjengelig i stor skala høsten 2022.[26] En analyse av mer enn 5000 bilder av bildegeneratoren Stable Diffusion fant at etnisitet- og kjønnsforskjeller trekkes til det ekstreme og forverrer skjevhetene som finnes i den virkelige verden.[27] Noe av bakgrunnen for dette kan trolig ha sammenheng med datagrunnlaget som modellene er trent på. Flere menn enn kvinner har tilgang til og er brukere av internett, og bruken er mer omfattende i den vestlige verden sammenlignet med Afrika og Asia.[28] Dette har konsekvenser fra hvem som legger igjen spor på nett, og vil igjen få betydning for datasettene som generativ AI baseres på.

Dersom slike systemer skal benyttes for å treffe avgjørelser om mennesker, er det avgjørende at datagrunnlaget er kvalitetssikret og godt begrunnet i de normative valgene som må tas ut ifra hvilken kontekst systemet skal operere i.

22 Riksrevisjonens rapport, «Bokontroll basert på maskinlæring» (sak 2019/01316), side 19.

23 Sluttrapport fra AHUS' sandkasseprosjekt 2022, tilgjengelig: https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/ferdige-prosjekter-og-rapporter/ahus-sluttrapport-ekg-ai/.

24 Se oversikt over ulike former for skjevheter i den vitenskapelig publikasjon «Bias og kvantitativ analyse innen velferd – opphav til skjevheter og relasjon til utfallsrettferdighet» av Andrea Marheim Storås, Robindra Prabhu, Hugo Lewi Hammer og Inga Strümke, tilgjengelig: https://www.idunn.no/doi/10.18261/tfv.25.3.3, hvor det diskuteres mulige løsninger på de ulike formene for skjevheter.

25 Ibid.

26 Herunder Midjourney, ChatGPT osv.

27 Bloombergs kartlegging av bildegeneratoren Stable Diffusion, tilgjengelig: https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

28 Se statistikk fra ITU, tilgjengelig: https://www.itu.int/itu-d/reports/statistics/2021/11/15/the-gender-digital-divide/.

Aktuelle problemstillinger som bør drøftes

Behov for data: Innsamling

a. Hvilke data trengs for å oppnå formålet med modellen?
b. Har virksomheten tilgang til dataene internt, eller må de skaffes fra eksterne kilder?[29]

Datakvalitet: Kartlegg datagrunnlaget

c. Er datagrunnlaget representativt sett opp mot modellens formål? Er noen grupper over- eller underrepresentert?
d. Hva kan være konsekvensen av manglende representasjon?
e. Kan noen persongrupper ha avvikende datamønstre i forhold til hva modellen beregner? Ta utgangspunkt i diskrimineringsgrunnlagene.

Case: AHUS' hjertesviktalgoritme

Utfordringen knyttet til datagrunnlaget var et av spørsmålene i forbindelse med AHUS' hjertesviktalgoritme som mottok vår veiledning i Datatilsynets sandkasse i 2022.[30]

Sammen med AHUS kartla vi risikoen for diskriminering i algoritmen, herunder risikoen for at algoritmen ville fungere bedre på noen grupper sammenlignet med andre. I den anledning undersøkte vi om treningsdataene var representative for den delen av befolkningen som mottar helsehjelp fra nettopp AHUS.

To grupper utmerket seg med hensyn til å ha andre symptomer ved hjertesvikt. Dette gjaldt kvinner og personer med et afrikansk eller asiatisk opphav. AHUS måtte vite om algoritmen har gode nok data til å lære å gjenkjenne symptombildet til kvinner versus menn og pasienter med ulikt etnisk opphav. Dersom det ikke tas hensyn til dette, kan man risikere at algoritmen gir falske negative svar til personer med innvandrerbakgrunn, og at pasienter mottar dårligere helsehjelp på grunn av sin etnisitet. Dette kan være diskriminerende.

29 Ved bruk av eksterne kilder krever en ytterligere årvåkenhet med hensyn til å kartlegge datakvaliteten.

30 Datatilsynets sluttrapport av februar 2023, tilgjengelig: https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/ferdige-prosjekter-og-rapporter/ahus-sluttrapport-ekg-ai/.

DEMO

3. Modellutvikling

Denne fasen handler om hvordan modellen skal utvikles for å oppnå det definerte formålet. Det avgjørende i denne fasen er at det modellen beregner, korresponderer med det man ønsker å oppnå.

Relevante diskrimineringsutfordringer

En typisk kilde til diskriminering handler om at det modellen beregner, ikke korresponderer med det egentlige formålet. Dette kan omtales som *måleskjevhet*.[31] Utfordringen kan oppstå dersom formålet med systemet ikke er direkte observerbart eller målbart. Dataene og variablene som blir benyttet i systemet, utgjør dermed en forenkling av det overordnede målet, og det kan oppstå fare for flere svakheter ved systemet som helhet, herunder en diskrimineringsrisiko.

Aktuelle problemstillinger som bør drøftes

a. Hva skal modellen beregne?
b. I hvilken grad samsvarer beregningen med det overordnede formålet?
c. Hvilke variabler baserer modellens beregning seg på, og hvorfor er disse relevante?
d. Dersom variablene kan knyttes til et diskrimineringsgrunnlag – kan saklig formål, nødvendighet og forholdsmessighet påvises?[32]
e. Er det tilstrekkelig med én modell, eller bør du utvikle flere modeller som kan sammenliknes med hensyn til eventuelle skjevheter og rettferdighet?

31 Obermeyer, Nissan, Stern , Eaneff, Bembeneck, Mullainathan i «Algorithmic Bias Playbook» av juni 2021, tilgjengelig: https://www.chicagobooth.edu/research/center-for-applied-artificial-intelligence/research/algorithmic-bias/playbook og «Bias og kvantitativ analyse innen velferd – opphav til skjevheter og relasjon til utfallsrettferdighet» av Andrea Marheim Storås, Robindra Prabhu, Hugo Lewi Hammer og Inga Strümke, tilgjengelig: https://www.idunn.no/doi/10.18261/tfv.25.3.3.

32 Se DEL 2, 3. spørsmål: Ulovlig.

Case: Amerikansk helsehjelpspredikator

Et illustrerende eksempel er en beslutningsstøttemodell benyttet i det amerikanske helsevesenet som skulle forutsi pasientenes behov for fremtidig helsehjelp.[33] Som en beregning av dette behovet tok modellen utgangspunkt i ulike pasientgrupper historiske utgifter til helsehjelp. Systemet tok imidlertid ikke høyde for ulike befolkningsgruppers sosioøkonomiske status og forutsetninger, og systemet antok derfor feilaktig at personer med afroamerikansk opphav hadde mindre behov for helsehjelp enn det som var tilfellet, sammenlignet med den hvite befolkningen.

Selv om norsk helsevesen er hjørnesteinen i velferdsstaten og er rigget svært annerledes enn det amerikanske, viser eksemplet betydningen av at systemets beregning korresponderer med modellens formål.

4. Testing av systemet

Denne fasen handler om hvordan systemet bør testes før det implementeres. Ettersom testingen må avgrenses i omfang og tid, er det vesentlig at systemet testes for relevante risikoer.

Relevante diskrimineringsutfordringer

Selv om systemet ikke benytter seg av personopplysninger som utgjør et diskrimineringsgrunnlag, så er det en risiko for at andre, tilsynelatende nøytrale opplysninger kan avdekke sammenhenger som sammenfaller med diskrimineringsgrunnlagene.[34]

Årsaken til dette er at maskinlæringsmodeller ofte er overlegne til å avdekke sammenhenger, men har begrenset evne til å skille mellom kausalitet og korrelasjoner. Det vil si sammenhenger som man kjenner årsakene til, og sammenhenger som man ikke kjenner årsakene til. Selv om man fjerner opplysninger om f.eks. kjønn eller etnisitet fra datagrunnlaget, evner maskinene å sammenstille mønstre i treningsdataene på en slik måte at den kan vektlegge opplysninger som likevel avdekker kjønn eller etnisitet.[35]

Forskning viser at det kan være nyttig å beholde personopplysningene som direkte fanges opp av diskrimineringsgrunnlagene for å kunne teste systemet og kartlegge om ulike persongrupper kommer dårligere ut enn andre.[36]

En særlig utfordring som må testingen av systemet må ta hensyn til, er det ovennevnte fenomenet *sammensatt diskriminering*.[37] Dersom systemet testes for forskjellsbehandling på gruppenivå helt overordnet (slik som kjønn), kan forskjellsbehandling på mer finmaskede gruppenivåer oversees. Kombinasjonene av diskrimineringsgrunnlag er så å si uendelige (f.eks. ulike aldersgrupper, kjønn, *ulike variasjoner* av funksjonsnedsettelser, ulike etnisiteter).

Dette demonstrerer betydningen av kunnskap om samfunnsforhold og ulike gruppers forutsetninger i den konteksten systemet skal fungere i, for å kunne teste systemet for sannsynlige svakheter.

33 Obermeyer, «Dissecting racial bias in an algorithm used to manage the health of populations», Science 366, no. 6464 (2019), s. 447–453, tilgjengelig: https://www.science.org/doi/10.1126/science.aax2342.

34 Se DEL 2, 2. spørsmål: Diskrimineringsgrunnlag.

35 Instituttet for menneskerettigheter i Nederland (College voor de Rechten van de Mens) er i ferd med å utvikle en oversikt over opplysninger som typisk sammenfaller med diskrimineringsgrunnlag og som bør vies særlig oppmerksomhet ved utvikling av ML-systemer.

36 Se blant annet: https://arxiv.org/abs/1104.3913.

37 Se DEL 2, 2. spørsmål: Diskrimineringsgrunnlag.

Aktuelle problemstillinger som bør drøftes

Gjennomfør testing av systemet:

a. Hvordan presterer modellen opp mot suksesskriteriene som er definert i fase 1? Herunder:
   i. Hvordan presterer systemets med hensyn til falske positive/falske negative for ulike grupper? Sammenlign resultatene for de ulike gruppene.
   ii. Er data for å kunne kontrollere for diskriminering for ulike grupper tilgjengelig?[38]
   iii. Hvem har ansvaret for oppfølging av modellens prestasjon på disse punktene?
b. Hvordan involveres representanter for de berørte i testfasen?

Korrelasjoner eller kausalitet:

Dokumenter bakgrunn for sammenhenger

c. Hva er de bakenforliggende årsakene til prediksjonene systemet gir?
d. Undersøk om sammenkobling av data kan utlede personopplysninger som kan knyttes til diskrimineringsgrunnlagene.
   i. Hvis ja – hva er begrunnelsen for dette?
   ii. Vurder saklig formål, nødvendighet og proporsjonalitet, jf. ldl. § 9

Case: Amazons rekrutteringsalgoritme

Amazons rekrutteringsalgoritme illustrerer problematikken som skisseres over.[39] Selskapet ønsket å effektivisere de interne ansettelsesprosessene ved å automatisere vurderingen av jobbsøknader ved hjelp av en maskinlæringsalgoritme. Etter noe tid viste det seg at algoritmen favoriserte mannlige søkere fremfor kvinnelige søkere, og verktøyet måtte avvikles.

I tillegg til at datagrunnlaget var kjønnsskjevt (altså en svakhet i treningsdataene, ref. risikoene omtalt under fase 2), viste det seg at algoritmen vektla ordvalg i søknadsbrevene på en diskriminerende måte. Algoritmen tok hensyn til bestemte nøkkelord og fraser som tidligere hadde vært assosiert med suksess i selskapet. Mange av disse ordene var mer typiske for mannlige søkere, som for eksempel uttrykk knyttet til en maskulin lederstil. Dette førte til at kvinnelige søkere ble ytterligere undervurdert. Selv om denne forskjellsbehandlingen i vektlegging av ord ikke retter seg spesifikt mot kvinner direkte, rammer den likevel kvinner indirekte.

38 Behovet for å teste systemet for diskriminering kan stå i konflikt med personvernforordningen (GDPR) art. 9 og hovedregelen om et forbud mot å samle inn særlige kategorier av personopplysninger. EU-kommisjonens forslag til ny forordning om kunstig intelligens (AI act) gir adgang til å behandle særlige kategorier av personopplysninger dersom det er strengt nødvendig for å overvåke, oppdage og korrigere algoritmeskjevhet, jf. artikkel 10 nr. 5. Dersom forslaget vedtas med nåværende ordlyd, kan det gi et lovgrunnlag for å kartlegge diskriminering i algoritmer ved å behandle særlig kategorier av personopplysninger.

39 Se blant annet: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

5. Implementering

Denne fasen handler om hvordan systemet blir tatt i bruk. Fasen forutsetter at testingen av systemet (fase 4) gir tilfredsstillende resultater.

Relevante diskrimineringsutfordringer

Dersom kartleggingen av spørsmålene i de forutgående fasene tilsier at systemet innebærer en risiko for at det vil fungere dårligere for enkelte grupper, eller behandler enkelte grupper strengere enn andre, og dette ikke kan justeres for i selve systemet, bør det undersøkes om det finnes en mulighet for å tilby en alternativ behandling som gir de aktuelle gruppene en likeverdig behandling.

Dersom det ML-systemet implementeres, er det avgjørende at det føres tilsyn med hvordan systemet presterer over tid. Denne teknologien er dynamisk og endrer seg basert på den erfaringen systemer gjør seg i den tiden den er tatt i bruk. Derfor kan en ikke-diskriminerende modell med tiden utvikle en diskriminerende slagside.

Aktuelle problemstillinger som bør drøftes

Supplementet til modellen:

a. Kan diskriminerende beregninger utført av systemet kompenseres for i møte med relevante persongrupper?

Kontroll:

b. Hvordan sikres reell menneskelig overprøving av modellens enkeltstående beslutninger?[40]
c. Hvordan sikres strukturell kontroll med systemet? Herunder:
   i. Kontrollmekanismer for at modellen gir like treffsikre beslutninger for alle grupper, og
   ii. Kontrollmekanismer for å påse at modellen ikke systematisk behandler enkelte grupper strengere.

Innsyn og kommunikasjon:

d. Hvem skal kunnskap om systemet og dets anvendelse være åpent og tilgjengelig for? Kartlegg hvilken informasjon de ulike gruppene av berørte trenger for å sikre tillit i samfunnet, og hvilke føringer det legger for hvordan informasjonen tilgjengeliggjøres.
e. Hvordan får de berørte av systemet ivaretatt sine interesser? Herunder:
   i. Nødvendig innsikt i hvordan modellen fungerer?
   ii. Hvilke diskrimineringsvurderinger som er gjort?
   iii. Finnes det reelle klagemuligheter?

Evaluering:

f. Definer en evalueringsstrategi (kontinuerlig eller periodevis), og involver gjerne eksterne fagkyndige, og interesseorganisasjoner som kan representere de berørte.
g. Hvordan ville systemet ha fungert med en alternativ modell, rettferdighetsdefinisjon eller algoritme?
h. Basert på evalueringen, bør systemet brukes videre, justeres eller avsluttes?

Case: AHUS' hjertesviktalgoritme

Et eksempel på dette er hjertesviktalgoritmen ved AHUS, nevnt over. AHUS hadde oversikt over pasientenes biologiske kjønn, men utfordringen var manglende data om pasientenes etniske opphav. Derfor var det heller ikke mulig å kontrollere hvorvidt algoritmen hadde mindre presise prediksjoner for pasientgrupper med etnisk minoritetsbakgrunn.

Ombudets vurdering var at AHUS likevel kunne kompensere for algoritmeskjevhet ved å supplere med andre medisinske metoder i undersøkelser av de pasientgruppene som algoritmen risikerer å ikke gi like gode prediksjoner. For å sikre like god behandling til alle pasienter må tjenesteyteren vite hvilke pasientgrupper algoritmen er mindre presis for, og iverksette tiltak slik at disse pasientene får et like godt tilbud som andre.

40 En typisk risiko i denne forbindelsen er at den menneskelige overprøvingen ikke er reell, fordi saksbehandlerens vurdering vil farges av ML-systemets anbefaling.

Case: Citybank Europe[41]

Et nederlandsk cateringsselskap med et syrisk stedsnavn i firmanavnet klaget inn Citibank Europe med påstand om etnisk diskriminering i forbindelse med en pengetransaksjon. Bankens screeningsystem blokkerte transaksjonen, og banken stilte krav om utvidet kontroll da selskapets navn medførte mistanke om terrorfinansiering.

Instituttet for menneskerettigheter i Nederland behandlet saken og kom til at bankens utvidede kontroll var lovlig forskjellsbehandling, men at klageren ble indirekte diskriminert fordi banken unnlot å behandle klagen fra cateringsselskapet.

Saken demonstrerer viktigheten av åpenhet om ML-systemene som er i bruk, samt å gi de berørte en reell klagemulighet. Dette bør inngå i en helhetlig plan for å sikre at de berørte er satt i stand til å ivareta sine rettigheter.

41 Avgjørelse fra Instituttet for menneskerettigheter i Nederland (College voor de Rechten van de Mens), Domsnummer 2023-29, av 28. februar 2023, tilgjengelig: https://oordelen.mensenrechten.nl/oordeel/2023-29.

Etterord

Dette er Likestillings- og diskrimineringsombudets første versjon av konseptet «Innebygd diskrimineringsvern». Enhver virksomhet som har spørsmål om forebygging av diskriminering eller hvordan de kan fremme likestilling ved utvikling eller bruk av kunstig intelligens, er velkommen til å ta kontakt med oss for individuell veiledning. Ombudet vil teste ut veilederen og gjøre justeringer i henhold til de erfaringene vi får i samhandling med de relevante fagmiljøene.

Å utføre risikovurderinger, slik denne veilederen legger opp til, er en konstruktiv måte å tilnærme seg de utfordringene ny teknologi kan innebære. Dersom EUs forordning om kunstig intelligens stiller krav til menneskerettslige konsekvensvurderinger ved bruk av høyrisikosystemer, kan slike risikovurderinger bli obligatoriske for virksomheter som tar teknologien i bruk.

Veilederen er blant annet inspirert av «Non-discrimination by design»,[42] Fundamental Rights and Algorithm Impact Assessment (FRAIA)[43] og «Promoting equality in the use of Artificial Intelligence – an assessment framework for non-discriminatory AI».[44]

Takk til Inga Strümke (NTNU), Helga Brøgger (DNV), Robindra Prabhu (NAV), Iris Bore (NAV), Rita Gyland (NAV), Jacob Sjødin (NAV), professor Dag Elgesem (UiB) og Vera Sofie Borgen Skjetne (BufDir), som har gitt innspill i prosessen med å utvikle denne veilederen.

42 Nederlandsk håndbok av Sloot, Keymolen og Noorman ved Tilburg universitet, The Netherlands Institute for Human Rights, Weerts, Wagensveld, Visser, tilgjengelig: https://www.tilburguniversity.edu/sites/default/files/download/04%20handbook%20non-discrimination%20by%20design%28ENG%29.pdf

43 Nederlandsk verktøy for menneskerettslig konsekvensvurdering av professor Gerards, Schäfer, Vankan og Muis ved Utrecht universitet, tilgjengelig: https://www.government.nl/documents/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms

44 Finsk rammeverk for vurdering av diskrimineringsrisiko av Ojanen, Björk, Helsinki, tilgjengelig: https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/164290/25-2022-Promoting%20equality%20in%20the%20use%20of%20Artificial%20Intelligence-an%20assessment%20framework%20for%20non-discriminatory%20AI.pdf?sequence=7&isAllowed=y

Likestillings- og diskrimineringsombudet

Vi hjelper deg

Gratis juridisk veiledning
Finn oss på nett: www.ldo.no
Ring oss: 23 15 73 00

facebook.com/mittOmbud

twitter.com/mittOmbud

instagram.com/mittOmbud