Dataetisk Råd

TECHSPLAINER

Hvad er algoritmisk bias?

Dataetisk Råd er en offentlig myndighed, som rådgiver om etiske udfordringer ved nye digitale teknologier.

https://dataetiskraad.dk/

Illustrationer er genereret med AI.

Dataetisk Råd, København 2026.

Denne techsplainer er udarbejdet i samarbejde med professor Kasper Lippert-Rasmussen og Center for Eksperimentel-Filosofiske Studier af Diskrimination (CEPDISC), Århus Universitet.

Hvad er algoritmisk bias?

En offentlig myndighed bruger en algoritme til at hjælpe med at vurdere hvilke socialt udsatte borgere, som har mest brug for en særlig indsats. Algoritmen vurderer oftere, at borgere med minoritetsbaggrund er særligt udsatte, end at borgere med majoritetsbaggrund er det.

Et hospital bruger kunstig intelligens til hjælpe med at diagnosticere kræftknuder. Den kunstige intelligens skelner mellem om patienten er mand eller kvinde, når den skal vurdere, om en knude er ondartet.

En virksomhed bruger en algoritme til at grovsortere jobansøgere. Algoritmen er mere præcis, når den vurderer ældre ansøgere, end når den vurderer yngre ansøgere.

Udvikling og udbredelse af nye digitale teknologier går stærkt. Vi bruger stadigt flere algoritmer og stadigt mere kunstig intelligens (AI).[1] Systemerne hjælper med at løse opgaver effektivt, vurdere svære forhold, og træffe vanskelige beslutninger. Men de rejser også komplekse etiske udfordringer.

Eksemplerne med udsatte borgere, patienter og jobansøgere ovenfor rejser en af de centrale udfordringer: risikoen for algoritmisk bias. Men hvad er algoritmisk bias? Og hvilke muligheder er der for at undgå det?

I denne techsplainer præsenterer Dataetisk Råd algoritmisk bias som begreb. Techsplaineren forklarer også hvordan algoritmisk bias kan opstå, introducerer nogle af de måder, man kan begrænse det, og diskuterer hvilke grupper, som kan blive udsat for algoritmisk bias.

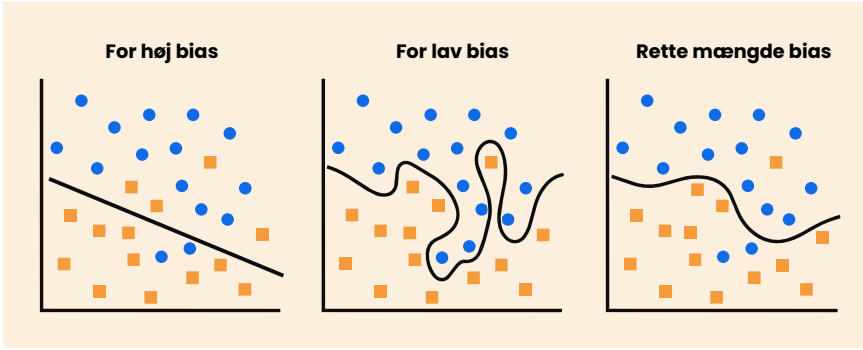
[1] AI kan forstås på forskellige måder. I denne sammenhæng betyder AI et computersystem, som er i stand til at løse en kompleks opgave, der normalt ville kræve menneskelig intelligens. Se Dataetisk Råds techsplainer "Hvad er generativ AI?" En algoritme er et sæt matematiske operationer, som et computersystem kan udføre for at løse en opgave. Alle former for AI bruger algoritmer, men en algoritme er ikke nødvendigvis en AI.

Hvad er bias?

Der findes to forskellige opfattelser af bias i debatten om algoritmisk bias.[2] Den første forstår blot bias som en tendens til systematisk at behandle eller vurdere noget på en bestemt måde. I denne forstand er bias i en AI eller algoritme blot den måde, teknologien anvender generelle regler til at behandle specifikke data.[3] Når man forstår bias på denne neutrale måde, kan en AI eller algoritme have både for lav og for høj bias.[4]

Bias (neutralt)

En tendens til systematisk at behandle eller vurdere noget på en bestemt måde.



Figur 1
En AI skal forsøge at skelne mellem de blå og orange datapunkter. En model med for høj bias bruger en simpel regel til at trække en ret linje. Det fører til at mange datapunkter lander på den forkerte side af linjen. En model med for lav bias bruger meget komplekse regler til at skelne. Det giver AI'en en høj risiko for at begå fejl, når den skal bruges på nye eksempler, fordi den ikke har lært hvordan datapunkterne generelt er forskellige. En model med den rette mængde bias bruger en tilpas kompleks regel til at skelne, som har stor sandsynlighed for også at ramme rigtigt på nye datapunkter.

[2] For et overblik over forskellige betydninger af bias, se Johnson 2024.

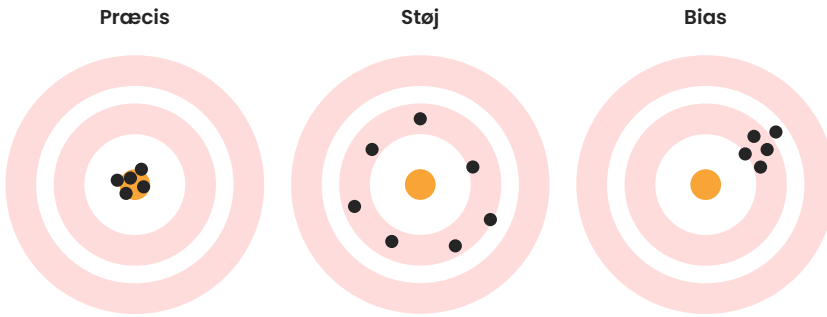
[3] Se eksempelvis Mitchell 1980; Gordon & Desjardins 1995.

[4] I faglitteraturen om maskinlæring kaldes denne balance for bias/varians-afvejningen. For lav bias gør modellen overtilpasset (eng. "overfitted"), mens for høj bias gør modellen undertilpasset (eng. "underfitted").

I almindelig tale, og i social- og kognitionspsykologien, er bias typisk mere negativt ladet. Her forstås bias som noget der forstyrrer vores måde at behandle eller vurdere ting, og som vi derfor bør forsøge at undgå.[5] I denne forstand betyder bias en tendens til systematisk at behandle eller vurdere noget på en måde, vi opfatter som problematisk. For eksempel har en almindelig terning en bias, hvis den oftere slår seks. Og en arbejdsgiver har en bias, hvis vedkommende systematisk lægger større vægt på mandlige ansøgeres kvalifikationer end på kvindelige ansøgeres kvalifikationer.

Bias (kritisk)

En problematisk tendens til systematisk at behandle eller vurdere noget.



Figur 2
Bias er systematiske afvigelser. En præcis model rammer tæt på rigtigt. En uprækis model rammer tilfældigt ved siden af – det kaldes "støj". En uprækis model med bias rammer systematisk ved siden af på en bestemt måde.

[5] Den danske ordbog definerer eksempelvis bias som: "[En] misvisende fremstilling af undersøgelsesresultater, målelige størrelser el.lign. som især skyldes metodiske fejl eller ubevidste præferencer", og "skævhed eller misforhold der skyldes forudfattede meninger og forestillinger". Se "Bias", Den danske ordbog (2023). For klassiske eksempler på den social- og kognitionspsykologiske forståelse af bias, se Kahneman & Tversky 1974; Samuelson & Zeckhauser 1988; Nickerson 1998; Bacchini & Lorusso 2019.

Hvad er algoritmisk bias?

Algoritmer og AI har ikke bias på samme måde som mennesker. En AI kan eksempelvis ikke have fordomme om et køn eller en etnisk gruppe. Ikke desto mindre har der det seneste årti været et intenst fokus på såkaldt "algoritmisk bias".[6] Ligesom bias mere generelt, kan algoritmisk bias i en kritisk forstand forstås som en AI eller algoritmes tendens til systematisk at behandle eller vurdere data på en måde, vi opfatter som problematisk.

Diskussionen om algoritmisk bias har især drejet sig om AI og algoritmer, som systematisk behandler visse *grupper* anderledes. For eksempel en algoritme, som behandler mænd anderledes end kvinder, unge anderledes end ældre, eller personer med mørk hud anderledes end personer med lys hud. Man taler i disse tilfælde ofte om, at algoritmen har bias *mod* en gruppe.

Algoritmisk bias (mod en gruppe)

En AI eller algoritmes problematiske tendens til systematisk at behandle eller vurdere personer fra én gruppe anderledes end personer fra andre grupper.

[6] For overblik, se: Carey & Wu 2023; Mitchell, et al. 2021; Heidari, et al. 2019; Barocas, et al. 2019; Chouldechova & Roth 2018; Fazelpour & Danks 2021. Algoritmisk bias er i dansk kontekst behandlet i eksempelvis Akhtar, et al. 2021; Dataetisk Råd 2024.

COMPAS

Meget af det seneste årtis interesse for algoritmisk bias udspringer af debatten om brugen af COMPAS i det amerikanske retsvæsen. COMPAS er en algoritme, som vurderer en persons risiko for recidivisme, dvs. fornyet, fremtidig kriminalitet. Den anvendes i USA blandt andet i forbindelse med varetægtsfængsling og prøveløsladelse. I 2016 offentliggjorde tænketanken ProPublica en analyse som hævdede, at COMPAS havde algoritmisk bias mod sorte amerikanere. (Larson et al. 2016; Angwin et al. 2016) ProPublica præsenterede dokumentation for, at COMPAS havde to tendenser. I gruppen af personer, som ikke efterfølgende begik en forbrydelse, blev sorte amerikanere oftere (fejl)vurderet til at have en høj risiko end hvide amerikanere. Og i gruppen af personer som efterfølgende faktisk begik en forbrydelse, blev hvide amerikanere oftere (fejl)vurderet til at have lav risiko end sorte amerikanere. Udvikleren af COMPAS forsvarede sig ved at pege på, at algoritmen var lige præcis: de personer, som fik samme risikoscore, havde samme sandsynlighed for at begå en ny forbrydelse, uanset om de var sorte eller hvide. (Dieterich et al. 2016) På paradoksal vis er begge påstande rigtige: COMPAS havde bias i den ene forstand, men ikke i den anden.

Fire slags algoritmisk bias

En algoritme eller AI har algoritmisk bias, når den systematisk vurderer eller behandler personer fra forskellige grupper forskelligt. Men hvad vil det sige, at vurdere eller behandle grupper forskelligt?[7] Der findes fire forskellige måder at forstå algoritmisk bias:

- Algoritmisk bias på grund af brug af data om grupper,
- Algoritmisk bias på grund af forskelle i vurderinger,
- Algoritmisk bias på grund af forskelle i præcision, og
- Algoritmisk bias på grund af forskelle i typen af fejl.

[7] Se eksempelvis Hellman 2020; Kleinberg, et al. 2019; Barocas & Selbst 2016; Dwork, et al. 2011; Friedler, et al. 2019; Barocas, et al. 2019; Chouldechova & Roth 2018; Hardt, et al. 2016; Corbett-Davies & Goel 2018. Et nyligt overblik findes i Carey & Wu 2023.

Brug af data om grupper

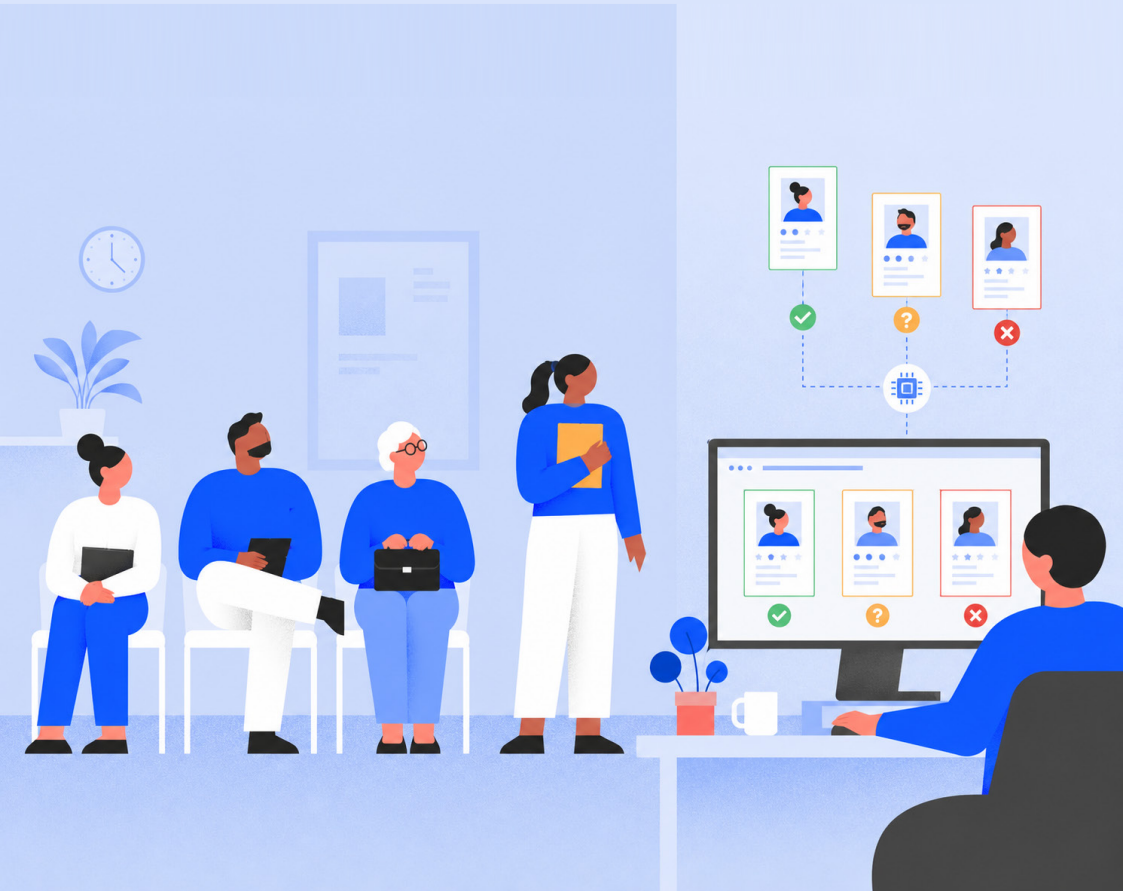
Algoritmer og AI behandler data. Ofte handler disse data om mennesker. En algoritme kan systematisk behandle personer fra en bestemt gruppe anderledes, ved ganske enkelt at bruge data om medlemskab af denne gruppe, når den skal lave en vurdering eller skabe indhold.[8]

Det er imidlertid ikke klart, at det altid udgør en form for algoritmisk bias, når en AI eller algoritme bruger information om medlemskab af en gruppe. For eksempel bliver det sjældent beskrevet som algoritmisk bias, hvis en AI bruger data om en patients køn til mere præcist at diagnosticere en sygdom, eller hvis en algoritme bruger information om en seers alder til at foreslå egnet indhold på en streamingtjeneste. Det skyldes måske at vi ikke opfatter dette som problematiske måder, at behandle grupperne forskelligt.

[8] Denne form for bias forbindes i faglitteraturen med idealet om "rimelighed gennem uvidenhed" (eng. "fairness through unawareness"). Kritikere har argumenteret for, at idealet er utilstrækkeligt, fordi en AI eller algoritme ofte kan have samme diskriminerende effekt uden brug af disse data, og kontraproduktivt, fordi det ofte vil være nødvendigt at bruge sådanne data netop for at modvirke andre former for algoritmisk bias. Se Dwork, et al. 2011; Barocas, et al. 2019; Lipton, et al. 2018.

Vurdering af risiko for langtidsledighed

I 2017 lancerede Styrelsen for Arbejdsmarked og Rekruttering en opdateret udgave af en algoritme, til at vurdere nylediges risiko for langtidsledighed. (MPLOY 2018; Moreau & Kulager 2021; Akhtar et al. 2021) Algoritmen blev kritiseret da det viste sig, at den blandt andet brugte data om lediges "ikke-vestlige herkomst" til at vurderer risikoen. (Bostrup 2021) Ligebehandlingsnævnet nåede i 2022 frem til, at der ikke var tale om ulovlig diskrimination. (Ligebehandlingsnævnets afgørelse om Etnisk oprindelse 2022). Ikke desto mindre blev brugen af data om "ikke-vestlig herkomst" af nogle opfattet som en form for algoritmisk bias.



Forskelle i vurdering

En AI eller algoritme kan godt behandle grupper forskelligt, selvom den ikke bruger data om personers medlemskab af grupperne. Det kan være tilfældet, hvis den alligevel har tendens til at *vurdere* personer forskelligt. Den kan for eksempel ofte vurdere personer fra én gruppe positivt, mens den omvendt ofte vurderer personer fra en anden gruppe negativt.[9] Sådanne forskelle er ofte i fokus i debatter om bias og forskelsbehandling, eksempelvis når vi taler om, at topposter i politik og erhvervsliv er ulige fordelt mellem mænd og kvinder.

Det er imidlertid ikke klart, at det altid udgør en form for algoritmisk bias, når en AI eller algoritme vurderer personer forskelligt.[10] En algoritme kunne eksempelvis vurdere at mænd, som har begået en forbrydelse, oftere har høj risiko for at begå ny kriminalitet, end kvinder, som har begået en forbrydelse. Hvis forskellen skyldes, at mænd oftere begår ny kriminalitet, kan det lyde underligt at sige, at der er tale om en algoritmisk bias. Igen kan det skyldes, at vi ikke opfatter denne forskel i behandling som problematisk.

[9] Denne form for algoritmisk bias forbindes i faglitteraturen med idealet om statistisk uafhængighed og demografisk lighed (eng. "independence" og "demographic parity"). Se Dwork, et al. 2011; Kusner, et al. 2017; Barocas, et al. 2019. Vi fokuserer her og i resten af teksten for nemheds skyld på AI og algoritmer som laver binære klassifikationer. De samme pointer gør sig gældende, med mindre justeringer, for AI som genererer andre typer resultater, f.eks. regressionsmodeller og generativ AI.

[10] Se eksempelvis Barocas, et al. 2019.

Vurdering af ansøgeres kompetencer

Amazon udviklede i 2014 en algoritme, som skulle hjælpe med at prioritere ansøgere til stillinger i virksomhedens software-afdelinger. Modellen blev trænet til at genkende gode ansøgere ved at lære af data om tidligere ansøgere og ansatte. I 2018 kom det frem, at Amazon havde skrinlagt algoritmen, fordi den havde tendens til at prioritere mænd frem for kvinder. Algoritmen anvendte ikke information om køn, men blev trænet på data, som næsten udelukkende indeholdt mandlige ansatte. Den lærte derfor at lægge vægt på visse karakteristika, som var udbredte blandt mandlige ansøgere, men ikke i samme grad blandt kvindelige. (Dastin 2022; Adams-Prassl et al. 2023)

Forskelle i præcision

Den tredje måde en AI eller algoritme kan behandle personer forskelligt, er ved at være mere *præcis* for én gruppe, og mindre præcis for andre grupper. En AI, som skal forsøge at vurdere om en elev har brugt en chatbot til snyde med en skriftlig opgave, kan eksempelvis være mindre præcis når den vurderer tosprogede elever, end når den vurderer elever, der kun har Dansk som modersmål.[11]

Den mest enkle måde at måle en AI eller algoritmes præcision er *overordnet præcision*, hvor man blot tæller hvor mange af vurderingerne, som er korrekte. Men der findes andre måder at måle en AI eller algoritmes præcision.[12] Hvis en AI skal forsøge at opdage elevers snyd med chatbots, så kan man for eksempel spørge: hvor mange af de elever, som algoritmen tror har snydt, har faktisk snydt? Denne målestok kaldes for positiv prædiktiv værdi.

[11] Denne type forskel er veldokumenteret i andre lande. Se Waked, et al. 2024; Liang, et al. 2023.

[12] Se Barocas, et al. 2019; Carey & Wu 2023.

Figur 3: Ti elever tager en test. AI vurderer af fire af dem bruger generativ AI til at snyde. Tre af disse fire elever snyder faktisk, mens én er uskyldig. Den positive prædiktive værdi er 3/4 = 75%. To elever som snyder, bliver ikke opdaget. Den overordnede præcision er 7/10 = 70%.

Omvendt kan man spørge: hvor mange af de elever, der faktisk har snydt, bliver fanget af den kunstige intelligens? Denne målestok kaldes for sensitivitet.[13]

Figur 4: Ti elever tager en test. Af de fem som snyder, bliver tre fanget af AI. Den kunstige intelligens' sensitivitet er 3/5 = 60%. Den overordnede præcision er stadig 7/10 = 70%.

[13] Sensitivitet kaldes også ofte for "sand positiv raten" og "genkald" (eng. "sensitivity/true positive rate/recall").

Fordi der findes forskellige måder at måle præcision, kan der også være flere forskellige slags algoritmisk bias. Man kan mene, at en AI eller algoritme har bias, hvis det er mere sandsynligt at den har ret, når den vurderer at én gruppe elever har snydt, end når den vurderer at andre elever har snydt (positiv prædiktiv værdi). Men man kan også mene at den har bias, hvis den for én gruppe fanger flere af de elever, som faktisk har snydt (sensitivitet).[14]

Ansigtsgenkendelse

Et af de mest kendte eksempler på algoritmisk bias på grund af forskelle i præcision drejer sig om ansigtsgenkendelse. Ansigtsgenkendelse anvendes blandt andet i moderne kameraer til at hjælpe med at fokusere på ansigter, på mange medieplatforme til automatisk at beskære billeder, og til at bekræfte en persons identitet ved adgangskontrol. Studier viser, at ansigtsgenkendelse ofte virker bedre for personer med lys hud end personer med mørk hud, ligesom der også kan være forskelle på tværs af køn og alder. (Grothar et al. 2019; Khalil et al. 2020; Robinson et al. 2020; Yucer et al. 2024; Vincent 2021) Det betyder, at et kamera kan have sværere ved at fokusere på nogle ansigter, at medieplatforme kan komme til at beskære visse personers ansigter på billeder, og at nogle grupper oftere oplever at blive fejlidentificeret (se også Dataetisk Råds techsplainer "Hvad er ansigtsgenkendelse?").

[14] Bias i form af forskelle i positiv prædiktiv værdi forbindes i faglitteraturen med idealet om tilstrækkelighed og vurderingslighed (eng. "sufficiency" og "predictive parity"). Se Barocas, et al. 2019; Carey & Wu 2023. Tilsvarende forbindes bias i form af forskelle i sensitivitet med idealet om adskillelse og lige odds (eng. "separation" og "equalised odds"). Se Barocas, et al. 2019; Carey & Wu 2023.

Forskelle i fejltyper

Den fjerde måde en AI eller algoritme kan behandle personer forskelligt, er ved at begå *forskellige slags* fejl. Det mest berømte eksempel er COMPAS (se ovenfor), som ifølge ProPublica havde tendens til ofte at *undervurdere* farlige hvide borgeres risiko for at begå ny kriminalitet, samtidig med at den *overvurderede* ufarlige sorte borgeres risiko for at begå ny kriminalitet. Selvom COMPAS begik fejl i begge tilfælde opfattede mange dette som en algoritmisk bias mod sorte amerikanere, fordi den ene type fejl var en fordel, mens den anden type fejl var en ulempe.

Ligesom vi har set i andre tilfælde, er det ikke klart, at det altid udgør en algoritmisk bias, når en AI eller algoritme begår forskellige typer fejl. Et eksempel kunne være et hospital, som bruger en algoritme, der har tendens til at overdiagnosticere ældre patienter og underdiagnosticere yngre patienter. Hvis det er dårligt både for en rask patient at blive fejldiagnosticeret og behandlet, og for en syg patient at blive fejldiagnosticeret og ikke behandlet, kan det være svært at sige, hvilken af de to grupper algoritmen skulle have bias imod. Hvis man afviser at der er tale om en algoritmisk bias, kan det måske igen skyldes, at man ikke opfatter forskellen i hvilken type fejl algoritmen begår som problematisk.

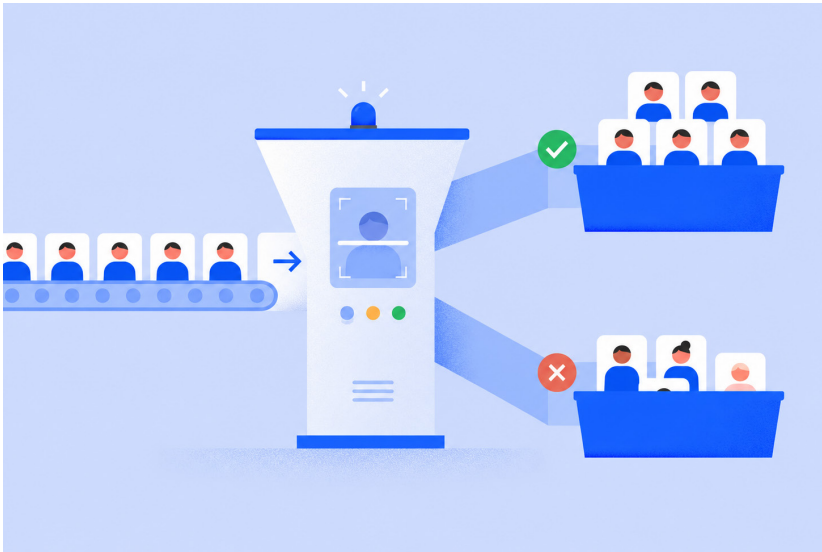
Hvor kommer algoritmisk bias fra?

Algoritmisk bias kan opstå på flere måder:

- **Data om grupper:** En AI eller algoritme bruger data om medlemskab af en relevant gruppe
- **Modellens type:** En AI eller algoritme behandler data på en måde, som virker bedre for visse grupper
- **Forskelle i data:** En AI eller algoritme bruger data som behandler grupper forskelligt eller viser forskelle på grupper.

Bias fra data om grupper

Selve det at bruge data om medlemskab af en gruppe, for eksempel oplysninger om en person køn, alder eller etnicitet, kan i sig selv opfattes som en form for algoritmisk bias (se ovenfor om brug af data om grupper). Men brug af data om medlemskab af en gruppe kan også påvirke for eksempel hvordan personer bliver vurderet, og hvor præcis en vurdering er. En AI kan for eksempel vurdere patienters behov på et hospital. Hvis der er statistisk forskel på yngre og ældre patienter, så vil brug af data om patienters alder typisk føre til, at AI'en behandler de to grupper forskelligt.



Bias fra modellens type

En AI eller algoritme kan også få algoritmisk bias ved at bruge en matematisk model, som virker bedre for nogle grupper end for andre. En AI eller algoritme bruger en matematisk model til at finde mønstre i data, når den skal løse en opgave. Forskellige slags AI og algoritmer tegner disse mønstre på forskellige måder. I nogle tilfælde er visse måder at finde mønstre bedre egnet til at behandle én gruppe, og mindre egnet til at behandle en anden. En algoritme kan for eksempel bruge en simpel matematisk regel, som ikke tager højde for, at en minoritetsgruppe er statistisk anderledes end gennemsnittet. Derved kan den komme til at vurdere grupperne forskelligt eller være mindre præcis for den ene gruppe.

Bias fra forskelle i data

Endelig kan en AI eller algoritme få algoritmisk bias på grund af forskelle i de data, den trækker på. Forskelle i data kan skabe bias både i udviklingen af en AI eller algoritme, og når den efterfølgende anvendes. Hvis der er forskelle i de data, som bruges til at udvikle en AI eller algoritme, så trænes den til at behandle grupperne forskelligt uanset hvilke data den anvender. Hvis der er forskelle i de data, en AI eller algoritme fodres med når den anvendes, så kan den få bias i sin behandling, uanset hvordan den er udviklet.

At der er forskelle i data kan betyde, at data om én gruppe har *lavere kvalitet* end data om andre grupper, at en gruppe *optræder mindre ofte* (eller slet ikke), eller at data *viser forskelle* på grupperne.

Hvis data om én gruppe har lavere kvalitet end data om andre grupper, vil en AI eller algoritme typisk blive mindre præcis, når den behandler personer fra denne gruppe. Forskelle i kvaliteten af data kan skyldes bias i den situation data handler om. Hvis for eksempel lærere har bias, og derfor giver lavere karakterer til kvindelige elever, så vil data om kvindelige elever systematisk undervurdere deres evner.[15] Forskelle i kvaliteten af data kan også skyldes bias i registreringen af data. Hvis for eksempel sundhedspersonale er mindre omhyggelige, når de noterer data om mandlige patienter, så

[15] Et meget omdiskuteret amerikansk eksempel på bias i de situationer som data beskriver er politiets overpatruljering af kvarterer med en stor andel af etniske minoriteter, hvilket fører til at disse minoriteter overrepræsenteres i statistikker om kriminalitet. Se: Ferguson 2017; Ensign, et al. 2017.

bliver mandlige patienters journaler mindre præcise end kvindelige patienters journaler. En AI eller algoritme, som bruger disse data, får sværere ved at vurdere mandlige patienter. I disse tilfælde reproducerer algoritmisk bias effekten af eksisterende bias.

Visse grupper kan også optræde mindre end andre grupper i data. Et datasæt om medarbejdere kan for eksempel have mange eksempler på mandlige ansatte, men få eksempler på kvindelige ansatte.[16] Faglitteraturen taler om, at grupper ofte ikke er lige godt *repræsenteret* i data. Hvis nogle grupper er mindre godt repræsenteret i data, så kan en AI eller algoritme overse relevante forskelle på grupperne, og behandle nogle grupper anderledes eller mindre præcist.[17]

Endelig kan data vise forskelle på grupper. Hvis data viser, at der er forskel på to grupper, så vil en AI eller algoritme typisk få tendens til at behandle de to grupper forskelligt. Data som viser forskelle på grupper kan være både faktuelle og fejlagtige. I det første tilfælde afspejler forskelle i data at der faktisk er forskel på de to grupper. I det andet tilfælde skyldes forskellen ofte bias i de situationer som data beskriver, eller bias i registrering af data (se ovenfor).

Det er vigtigt at være opmærksom på, at forskelle i data kan føre til algoritmisk bias, selvom informationer om medlemskab af grupper ikke direkte findes i data. For eksempel kunne data om kvinders uddannelse, indkomst, eller sundhed være mindre præcise end tilsvarende data om mænd, selvom data om køn ikke optræder i datasættet. Et datasæt kan også vise forskelle på grupper selvom der ikke er nogen data, som direkte handler om grupperne. For eksempel hænger data om en persons højde typisk sammen med køn, fordi mænd i gennemsnit er højere end kvinder. Data som viser en sammenhæng mellem personers højde og for eksempel indkomst, sundhed eller uddannelse, viser derfor også indirekte en sammenhæng mellem køn og disse data. I et stort datasæt vil mange

[16] Amazons algoritme til prioritering af ansøgere, som måtte skrinlægges, da den udviste algoritmisk bias mod kvindelige ansøgere, er et ofte omtalt eksempel. Se Dastin 2022; Adams-Prassl, et al. 2023.

[17] Et meget omtalt eksempel på bias gennem forskelle i repræsentation er ansigtsgenkendelse, som især i 2010'erne ofte havde vanskeligt ved at genkende ansigter på personer fra etniske minoriteter fordi algoritmerne fortrinsvis var trænet på billeder af personer fra de grupper, som udgør den etniske majoritet i vestlige lande. Se Robinson, et al. 2020; Bacchini & Lorusso 2019; Grothar, et al. 2019.

forskellige data ofte tilsammen hænge tæt sammen med medlemskab af en gruppe.[18] Brug af data som viser sådanne forskelle kan således føre til at en AI eller algoritme systematisk behandler grupper forskelligt.

Hvordan undgår man algoritmisk bias?

I mange situationer er de fleste enige om, at vi bør undgå algoritmisk bias. Især det seneste årti har forskere derfor udviklet en række tekniske metoder til at fjerne eller begrænse det.[19] Disse metoder kan inddeles i tre overordnede typer:

- Ændringer i træningsdata[20]
- Justering af træningsprocessen[21]
- Indgreb i vurderinger eller genereret indhold[22]

En AI eller avanceret algoritme udvikles typisk med maskinlæring, hvor den lærer at løse sin opgave ved at analysere træningsdata. En oplagt måde at begrænse algoritmisk bias er derfor at ændre på træningsdata.

Man kan for eksempel undgå den form for bias som består i, at en AI eller algoritme bruger data om medlemskab af en gruppe, ved at "blinde" en AI eller algoritme under udviklingen.[23] Det betyder at udvikleren fjerner data om medlemskab af relevante grupper fra

[18] I faglitteraturen tales om, at informationer om gruppemedlemskab bliver redundant indkodet (eng. "redundant encoding"), ved at andre data gør det muligt at forudsige for eksempel køn, alder, eller etnicitet. Se Dwork, et al. 2011; Lipton, et al. 2018.

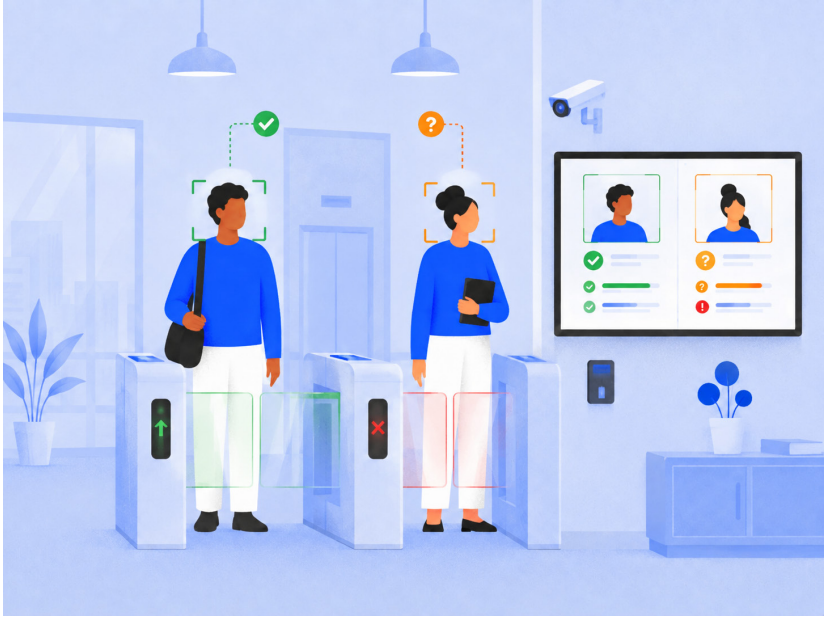
[19] For overblik, se: Friedler, et al. 2019; Chen, et al. 2023; Hort, et al. 2024.

[20] I faglitteraturen beskrives denne tilgang ofte som at forhindre bias gennem præ-processering (eng. "pre-processing").

[21] I faglitteraturen beskrives denne tilgang ofte som at forhindre bias gennem i-processering (eng. "in-processesing").

[22] I faglitteraturen beskrives denne tilgang ofte som at forhindre bias gennem post-processering (eng. "post-processesing").

[23] Dwork, et al. 2011; Kilbertus, et al. 2018.



træningsdata. Hvis en AI eller algoritme for eksempel ikke trænes på data om køn, kan den ikke lære at bruge køn til at vurdere personer eller generere indhold.

Mere avancerede ændringer af træningsdata kan begrænse bias i form af forskelle i vurderinger, præcision eller fejltyper. Udvikleren kan fjerne eller tilføje data, i nogle tilfælde ved ganske enkelt at duplikere visse data – for eksempel for at minoritetsgrupper kommer til at fylde mere i datasættet. Udvikleren kan også omskrive data – for eksempel ændre på nogle personers alder, så yngre og ældre kommer til at se mere ens ud i træningssættet. Eller udvikleren kan give nogle data større vægt – for eksempel således at en algoritme lærer mere af data om mænd end af data om kvinder.

En udvikler kan også begrænse algoritmisk bias ved at justere træningsprocessen. Når en AI eller algoritme skal lære at løse en opgave, bruger den en såkaldt tabsfunktion (eng. "loss function"). Tabsfunktionen er en matematisk målestok for hvor godt den løser opgaven. En udvikler kan give algoritmiske bias en negativ værdi i tabsfunktionen, og derved lære en AI eller algoritme at undgå dem.

Endelig kan en udvikler designe en AI eller algoritme, så den griber ind efter at den har forsøgt at løse sin opgave, men inden resultatet leveres til brugeren. For eksempel kan en AI eller algoritme bruge forskellige matematiske tærskler for forskellige grupper, eller

ændre visse resultater, så grupper behandles mere lige. Et almindeligt eksempel er indholdsfiltre på generativ AI. Mange populære chatbots er udstyret med digitale filtre, som vurderer det indhold, den generative AI skaber, før det vises til brugeren. Filtrene er designet til at opdage for eksempel racistisk eller sexistisk indhold. Hvis filteret opdager denne type indhold, så blokeres det, og chatbotten svarer i stedet for eksempel, at indholdet overtræder chatbottens retningslinjer.

Kan man fjerne alle slags algoritmisk bias?

Ved at ændre træningsdata, justere træningen, og gribe ind i resultatet er det ofte muligt at reducere eller undgå bestemte former for algoritmisk bias. Men kan man fjerne alle former for algoritmisk bias? Et overraskende resultat i den nyere forskning er, at dette ofte er omkostningsfuldt, vanskeligt eller endda i praksis umuligt.

En første udfordring er, at mange metoder til at begrænse bias reducerer en AI eller algoritmes præcision. I disse tilfælde har det en pris at undgå algoritmisk bias – AI'en eller algoritmen kommer til at lave flere fejl.[24]

En anden udfordring er, at de metoder man kan bruge til at undgå bias i form af forskelle i vurderinger, præcision og fejltyper, normalt forudsætter, at man anvender data om gruppemedlemskab. Det betyder, at man for at undgå disse tre former for bias kommer til at introducere bias i form af forskellig behandling af personer på baggrund af deres gruppemedlemskab.[25]

En tredje udfordring er at det i realistiske situationer ofte er sådan, at man ved at reducere nogle bias uundgåeligt forstærker andre bias. Denne udfordring er demonstreret i såkaldte umulighedsteoremer for algoritmisk bias.[26]

[24] Se Corbett-Davies, et al. 2017; Menon & Williamson 2018; Friedler, et al. 2019.

[25] Se Lipton, et al. 2018.

[26] To forskergrupper producerede i kølvandet på COMPAS-debatten uafhængigt matematiske beviser for, at det kun i helt særlige situationer kan lade sig gøre på samme tid at undgå bias i form af forskelle i positiv og negativ prædiktiv værdi, og bias i form af forskelle i sand positiv rate (sensitivitet) og sand negativ rate (specificitet). Se Kleinberg, et al. 2016; Chouldechova 2017; Mitchell, et al. 2021. Det er værd at bemærke, at de samme udfordringer gør sig gældende for menneskelige beslutningstagere.

I mange situationer er spørgsmålet derfor ikke hvordan man undgår algoritmisk bias, men *hvilke* former for algoritmisk bias man bør undgå, og hvilken pris man er villig til at betale, for at undgå dem.

Hvilke grupper kan blive udsat for algoritmisk bias?

Algoritmisk bias handler om, at AI og algoritmer behandler grupper forskelligt. Et centralt spørgsmål er derfor hvilke grupper, som kan blive udsat for algoritmisk bias? Både i faglitteraturen og den offentlige debat antages det ofte, at en AI eller algoritme kan have bias mod eksempelvis kvindelige medarbejdere, mod yngre patienter eller mod elever med minoritetsbaggrund. Omvendt antager de fleste, at den ikke kan have algoritmisk bias mod eksempelvis venstrehåndede medarbejdere, mod patienter som er håndboldfans, eller mod elever hvis navn begynder med bogstavet T. Men hvorfor det?

Et muligt svar er, at vores forståelse af algoritmisk bias trækker på begreber om diskrimination, som typisk afgrænses til bestemte grupper.[27] Nogle forskere har imidlertid argumenteret for, at AI og algoritmer også kan have bias mod nye og anderledes grupper, inklusive såkaldt "emergente grupper", som kun findes i den forstand at de er genstand for udbredt algoritmisk bias.[28]. Sådanne grupper kan være meningsløse for mennesker – det kan for eksempel være en gruppe, som har en bestemt kombination af mange forskellige slags uskyldig adfærd som brugere på internettet – men bør ifølge denne tankegang alligevel kunne kritisere en AI eller algoritme for at have bias imod dem.

[27] I faglitteraturen argumenterer nogle forskere, at diskrimination drejer sig om særligt socialt fremtrædende grupper, bl.a. fordi disse grupper kan blive genstand for systematisk forskelsbehandling. Se Lippert-Rasmussen 2013.

[28] Se Kuppler, et al. 2022; Zeiser 2025; Hu 2025.

Algoritmisk bias og dataetik

Mange opfatter algoritmisk bias som etisk problematisk. Hvis man forstår algoritmisk bias neutralt, er det et selvstændigt spørgsmål hvornår og hvorfor nogle former for algoritmisk bias, er etisk problematiske. Hvis man forstår algoritmisk bias kritisk, så er der kun tale om algoritmisk bias, når en AI eller algoritme behandler personer forskelligt på en måde, som er problematisk. I begge tilfælde er det et væsentligt spørgsmål, hvad det er som gør eller kan gøre algoritmisk bias etisk problematisk. *Hvorfor* er algoritmisk bias et problem?

Der findes flere forskellige svar på dette spørgsmål. Nogle forskere har argumenteret, at algoritmisk bias er etisk problematisk når og fordi det gør skade.[29] Andre hævder, at algoritmisk bias er etisk problematisk når det er uretfærdigt, for eksempel fordi det behandler personer urimeligt (eng. "unfair"), eller skaber ulighed.[30] En beslægtet pointe i faglitteraturen er, at det kan spille en rolle hvilke bias menneskelige beslutningstagere har – hvis alternativet til en AI eller algoritme med bias er menneskelige beslutninger med endnu større bias, kan man mene, at algoritmisk bias udgør en mindre udfordring.

Fordi der findes forskellige opfattelser og komplikationer fortjener det grundig overvejelse, hvordan man etisk bør vurdere algoritmisk bias, når man tager stilling til algoritmisk bias i en konkret kontekst.

[29] Se Heidari, et al. 2018; Hu & Chen 2020; Mittelstadt, et al. 2023.

[30] Se Heidari, et al. 2019; Eva 2022; Hedden 2021; Long 2021; Beigang 2023; Loi, et al. 2024.

Litteratur

Adams-Prassl, J., Binns, R. & Kelly-Lyth, A. (2023). "Directly Discriminatory Algorithms." *The Modern Law Review* **86**(1): 144-175 DOI: https://doi.org/10.1111/1468-2230.12759.

Akhtar, M., Thomsen, F. K., Jørgensen, R. F. & Boye Koch, P. (2021). *Når algoritmer sagsbehandler*, Institut for Menneskerettigheder. https://menneskeret.dk/files/media/document/Algoritmer_8.K.pdf.

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016) "Machine Bias." *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bacchini, F. & Lorusso, L. (2019). "Race, again: how face recognition technology reinforces racial discrimination." *Journal of Information, Communication and Ethics in Society* **17**(3): 321-335 DOI: 10.1108/JICES-05-2018-0050.

Barocas, S., Hardt, M. & Narayanan, A. (2019). *Fairness and machine-learning*. https://fairmlbook.org/.

Barocas, S. & Selbst, A. D. (2016). "Big Data's Disparate Impact." *California Law Review* **104**(3): 671-732 DOI: 10.2139/ssrn.2477899.

Beigang, F. (2023). "Reconciling Algorithmic Fairness Criteria." *Philosophy & Public Affairs* **51**(2): 166-190 DOI: https://doi.org/10.1111/papa.12233.

Bostrup, J. (2021). *Tellis oplevelse på jobcenteret fører nu til en optrapning i kampen mod diskriminerende algoritmer*. Politiken. https://politiken.dk/viden/Tech/art8140892/Tellis-oplevelse-p%C3%A5-jobcenteret-f%C3%B8rer-nu-til-en-optrapning-i-kampen-mod-diskriminerende-algoritmer.

Carey, A. N. & Wu, X. (2023). "The statistical fairness field guide: perspectives from social and formal sciences." *AI and Ethics* **3**(1): 1-23 DOI: 10.1007/s43681-022-00183-3.

Chen, J., Dong, H., Wang, X., Feng, F., Wang, M. & He, X. (2023). "Bias and debias in recommender system: A survey and future directions." *ACM Transactions on Information Systems* **41**(3): 1-39.

Chouldechova, A. (2017). "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." *Big Data* **5**(2): 153-163. https://ui.adsabs.harvard.edu/\#abs/2016arXiv161007524C.

Chouldechova, A. & Roth, A. (2018) "The Frontiers of Fairness in Machine Learning." *arXiv e-prints*. https://ui.adsabs.harvard.edu/\#abs/2018arXiv181008810C.

Corbett-Davies, S. & Goel, S. (2018). "The measure and mismeasure of fairness: A critical review of fair machine learning." *arXiv preprint arXiv:1808.00023*.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. & Huq, A. (2017). *Algorithmic decision making and the cost of fairness*. KDD '17.

Dastin, J. (2022). "Amazon scraps secret AI recruiting tool that showed bias against women." *Ethics of data and analytics*, Auerbach Publications**:** 296-299.

Dataetisk Råd (2024). *Dataetiske perspektiver ved brug af medarbejderdata* København. https://www.dataetiskraad.dk/media/x2fp2fip/dataetiske_perspektiver_medarbejderdata_v12_web.pdf.

Dieterich, W., Mendoza, C. & Brennan, T. (2016). *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*, NorthPointe.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O. & Zemel, R. (2011). "Fairness Through Awareness." *arXiv:1104.3913 [cs]*. http://arxiv.org/abs/1104.3913.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C. & Venkatasubramanian, S. (2017). *Runaway Feedback Loops in Predictive Policing*. 1st Conference on Fairness, Accountability and Transparency, arXiv. https://arxiv.org/abs/1706.09847.

Eva, B. (2022). "Algorithmic Fairness and Base Rate Tracking." *Philosophy & Public Affairs* **50**(2): 239-266 DOI: https://doi.org/10.1111/papa.12211.

Fazelpour, S. & Danks, D. (2021). "Algorithmic bias: Senses, sources, solutions." *Philosophy Compass* **16**(8): 1-16 DOI: https://doi.org/10.1111/phc3.12760.

Ferguson, A. G. (2017). *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*, NYU Press.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P. & Roth, D. (2019). *A comparative study of fairness-enhancing interventions in machine learning*. Proceedings of the Conference on Fairness, Accountability, and Transparency. Atlanta, GA, USA, Association for Computing Machinery**:** 329–338 DOI: 10.1145/3287560.3287589.

Gordon, D. F. & Desjardins, M. (1995). "Evaluation and Selection of Biases in Machine Learning." *Machine Learning* **20**(1): 5-22 DOI: 10.1023/A:1022630017346.

Grothar, P., Ngan, M. & Hanaoka, K. (2019). *Face Recognition Vendor Test (FRVT) - Part 3: Demographic Effects*. U.S. Department of Commerce, National Institute of Standards and Technology.

Hardt, M., Price, E. & Srebro, N. (2016). "Equality of Opportunity in Supervised Learning." *arXiv:1610.02413 [cs]*. http://arxiv.org/abs/1610.02413.

Hedden, B. (2021). "On statistical criteria of algorithmic fairness." *Philosophy & Public Affairs* **49**(2): 209-231 DOI: https://doi.org/10.1111/papa.12189.

Heidari, H., Ferrari, C., Gummadi, K. & Krause, A. (2018). "Fairness behind a veil of ignorance: A welfare analysis for automated decision making." *Advances in neural information processing systems* **31**.

Heidari, H., Loi, M., Gummadi, K. P. & Krause, A. (2019). *A moral framework for understanding fair ml through economic models of equality of opportunity*. Proceedings of the conference on fairness, accountability, and transparency.

Hellman, D. (2020). "Measuring algorithmic fairness." *Virginia Law Review* **106**(4): 811-866.

Hort, M., Chen, Z., Zhang, J. M., Harman, M. & Sarro, F. (2024). "Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey." *ACM J. Responsib. Comput.* **1**(2): Article 11 DOI: 10.1145/3631326.

Hu, L. (2025). "Does calibration mean what they say it means; or, the reference class problem rises again." *Philosophical Studies* **182**(5).

Hu, L. & Chen, Y. (2020). *Fair classification and social welfare*. Proceedings of the 2020 conference on fairness, accountability, and transparency.

Johnson, G. M. (2024). "Varieties of Bias." *Philosophy Compass*(7): e13011.

Kahneman, D. & Tversky, A. (1974). "Judgment under Uncertainty: Heuristics and Biases." *Science* **185**: 1124-1131.

Khalil, A., Ahmed, S. G., Khattak, A. M. & Al-Qirim, N. (2020). "Investigating Bias in Facial Analysis Systems: A Systematic Review." *IEEE Access* **8**: 130751-130761 DOI: 10.1109/ACCESS.2020.3006051.

Kilbertus, N., Gascón, A., Kusner, M. J., Veale, M., Gummadi, K. P. & Weller, A. (2018). "Blind Justice: Fairness with Encrypted Sensitive Attributes." *arXiv:1806.03281 [cs, stat]*. http://arxiv.org/abs/1806.03281.

Kleinberg, J., Ludwig, J., Mullainathan, S. & Sunstein, C. R. (2019) "Discrimination in the Age of Algorithms." *arXiv e-prints*. https://ui.adsabs.harvard.edu/\#abs/2019arXiv190203731K.

Kleinberg, J., Mullainathan, S. & Raghavan, M. (2016) "Inherent Trade-Offs in the Fair Determination of Risk Scores." *arXiv e-prints*. https://ui.adsabs.harvard.edu/\#abs/2016arXiv160905807K.

Kuppler, M., Kern, C., Bach, R. L. & Kreuter, F. (2022). "From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making." *Frontiers in Sociology* **Volume 7 - 2022** DOI: 10.3389/fsoc.2022.883999.

Kusner, M. J., Loftus, J. R., Russell, C. & Silva, R. (2017) "Counterfactual Fairness." *arXiv e-prints*. https://ui.adsabs.harvard.edu/\#abs/2017arXiv170306856K.

Larson, J., Mattu, S., Kirchner, L. & Angwin, J. (2016) "How We Analyzed the COMPAS Recidivism Algorithm." *ProPublica*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E. & Zou, J. (2023). "GPT detectors are biased against non-native English writers." *Patterns (N Y)* **4**(7): 100779 DOI: 10.1016/j.patter.2023.100779.

Ligebehandlingsnævnets afgørelse om Etnisk oprindelse (2022). Ligebehandlingsnævnet. https://www.retsinformation.dk/eli/retsinfo/2022/9518.

Lippert-Rasmussen, K. (2013). *Born Free and Equal? A Philosophical Inquiry Into the Nature of Discrimination*. Oxford, Oxford University Press.

Lipton, Z. C., Chouldechova, A. & McAuley, J. (2018). *Does mitigating ML's impact disparity require treatment disparity?* 32nd Conference on Neural Information Processing Systems.

Loi, M., Herlitz, A. & Heidari, H. (2024). "Fair equality of chances for prediction-based decisions." *Economics and Philosophy* **40**(3): 557-580 DOI: 10.1017/S0266267123000342.

Long, R. (2021). "Fairness in machine learning: Against false positive rate equality as a measure of fairness." *Journal of Moral Philosophy* **19**(1): 49-78.

Menon, A. K. & Williamson, R. C. (2018). *The cost of fairness in binary classification*. Conference on Fairness, accountability and transparency, PMLR.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A. & Lum, K. (2021). "Algorithmic Fairness: Choices, Assumptions, and Definitions." *Annual Review of Statistics and Its Application* **8**(1): 141-163 DOI: 10.1146/annurev-statistics-042720-125902.

Mitchell, T. M. (1980). *The need for biases in learning generalizations*. Rutgers CS tech report, Rutgers University. https://www.cs.cmu.edu/~tom/pubs/NeedForBias_1980.pdf.

Mittelstadt, B., Wachter, S. & Russell, C. (2023). "The Unfairness of Fair Machine Learning: Leveling Down and Strict Egalitarianism by Default." *Mich. Tech. L. Rev.* **30**: 1.

Moreau, T. & Kulager, F. (2021). *Vi har skilt jobcentrenes algoritme ad*. Zetland. https://www.zetland.dk/historie/sOMVZ7qG-aOz9m93B-a30b8.

MPLOY (2018). *Evaluering af projekt "Samtaler og indsats der modvirker langtidsledighed"*, Styrelsen for Arbejdsmarked og Rekruttering, . https://star.dk/media/8004/evaluering-af-projekt-samtaler-og-indsats-der-modvirker-langtidsledighed.pdf.

Nickerson, R. S. (1998). "Confirmation bias: A ubiquitous phenomenon in many guises." *Review of general psychology* **2**(2): 175-220.

Robinson, J. P., Livitz, G., Henon, Y., Qin, C., Fu, Y. & Timoner, S. (2020). *Face recognition: too bias, or not too bias?* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.

Samuelson, W. & Zeckhauser, R. (1988). "Status quo bias in decision making." *Journal of Risk and Uncertainty* **1**(1): 7-59 DOI: 10.1007/bf00055564.

Vincent, J. (2021) "Twitter's photo-cropping algorithm prefers young, beautiful, and light-skinned faces." *The Verge*. https://www.theverge.com/2021/8/10/22617972/twitter-photo-cropping-algorithm-ai-bias-bug-bounty-results.

Waked, A. N., Abdelsalem, H., Ashraf, M. W., Pilotti, M. & El Alaoui, K. (2024). "The Impact of Falsely Detecting AI-Generated Text on Academic Assessment." *International Society for Technology, Education, and Science*.

Yucer, S., Tektas, F., Moubayed, N. A. & Breckon, T. (2024). "Racial Bias within Face Recognition: A Survey." *ACM Comput. Surv.* **57**(4): Article 105 DOI: 10.1145/3705295.

Zeiser, J. (2025). "Emergent Discrimination: Should We Protect Algorithmic Groups?" *Journal of Applied Philosophy* **42**(3): 910-928 DOI: https://doi.org/10.1111/japp.12793.