Kursus i algoritmiske grundprincipper og etik i AI: fra teori til praksis

Værktøjer til synkrone sessioner

CU4 | Data-fairness og bias i AI

Understøttende PowerPoint-slides

Projektnummer | 2022-1-ES01-KA220-HED-000085257

INDHOLD

INDLEDNING

BILLEDKILDE | Genereret af DALL-E

Forslag

Start med en **interaktiv afstemning eller et spørgsmål** for at vurdere deltagernes umiddelbare forståelse eller personlige erfaringer med AI-fairness og bias. Det kan gøre introen mere engagerende og sætte scenen for, hvorfor emnet er relevant.

- Et eksempel: Hvad er dit kendskab til fairness og bias i AI?
  - Jeg kender det slet ikke.
  - Jeg har hørt om det, men jeg kender ikke detaljerne om det.
  - Jeg forstår det grundlæggende, men jeg vil gerne vide mere.
  - Jeg ved ret meget om det og har arbejdet med AI-systemer, der tager fairness og bias i betragtning.

I DENNE KOMPETENCEENHED FINDER DU FØLGENDE EMNER:

- Det grundlæggende vedrørende data-fairness og bias i AI.
- Forskellige bias, der kan påvirke AI-systemer.
- Hvordan man spotter og måler bias i data.
- Teknikker til at reducere bias og øge retfærdighed.
- Metrikker til at vurdere AI-retfærdighed.
- Bedste praksis for fair dataindsamling og -behandling.
- Nye tendenser inden for AI-fairness og bias i AI.

VED AFSLUTNINGEN AF KOMPETENCEENHEDEN, BØR DU VÆRE I STAND TIL AT:

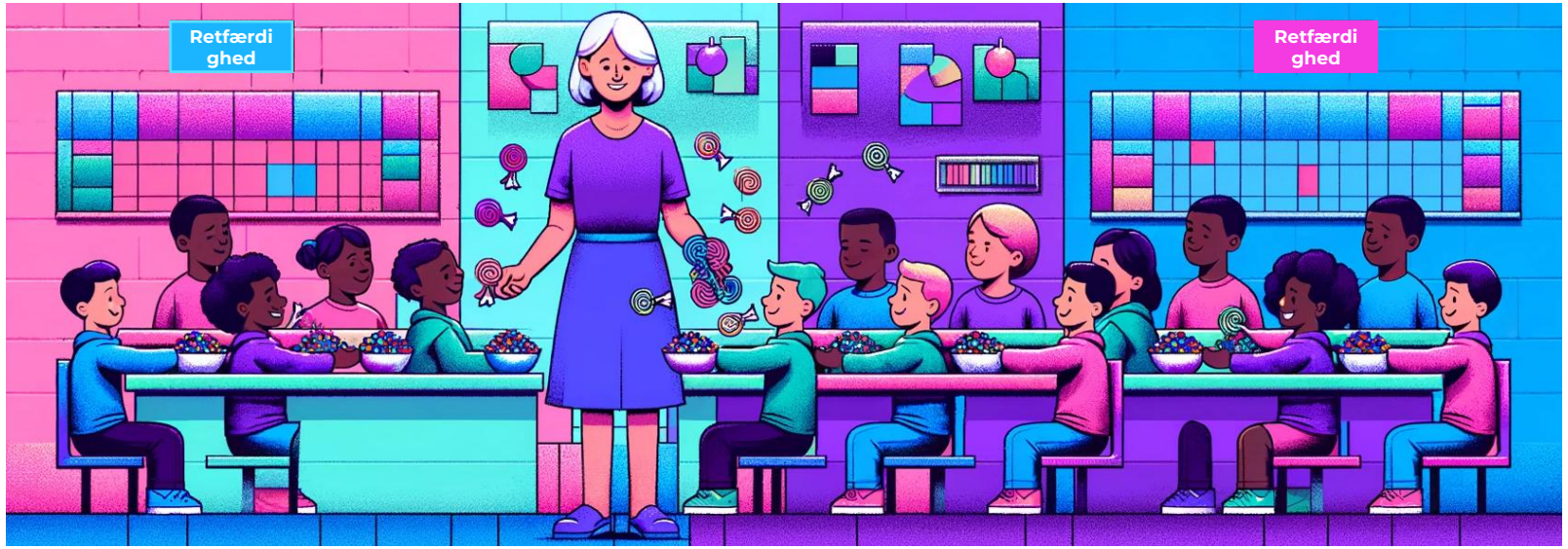
- Forklare de grundlæggende principper for data-fairness og betydningen af bias i AI.
- Identificere de forskellige former for bias, der kan påvirke AI-systemer.
- Vurdere AI-systemer ved hjælp af fairness-metrikker.
- Følge *best practice* for indsamling og forbehandling af data for at fremme retfærdighed.

INTRODUKTION TIL DATA-FAIRNESS OG BIAS

BILLEDKILDE | Genereret af DALL-E

Introduktion til data-fairness og bias

Fairness og bias – slikanalogien



BILLEDKILDE | Genereret af DALL-E

Oplev, hvordan retfærdighed og bias i AI-systemer kan påvirke vores daglige valg, ligesom en lærers regler for uddeling af slik kan præge en oplevelse i klasseværelset.

Lad os pakke begrebet AI-retfærdighed ud – et sødt eksempel ad gangen!

Introduktion til data-fairness og bias

Fairness og bias – slik-analogien

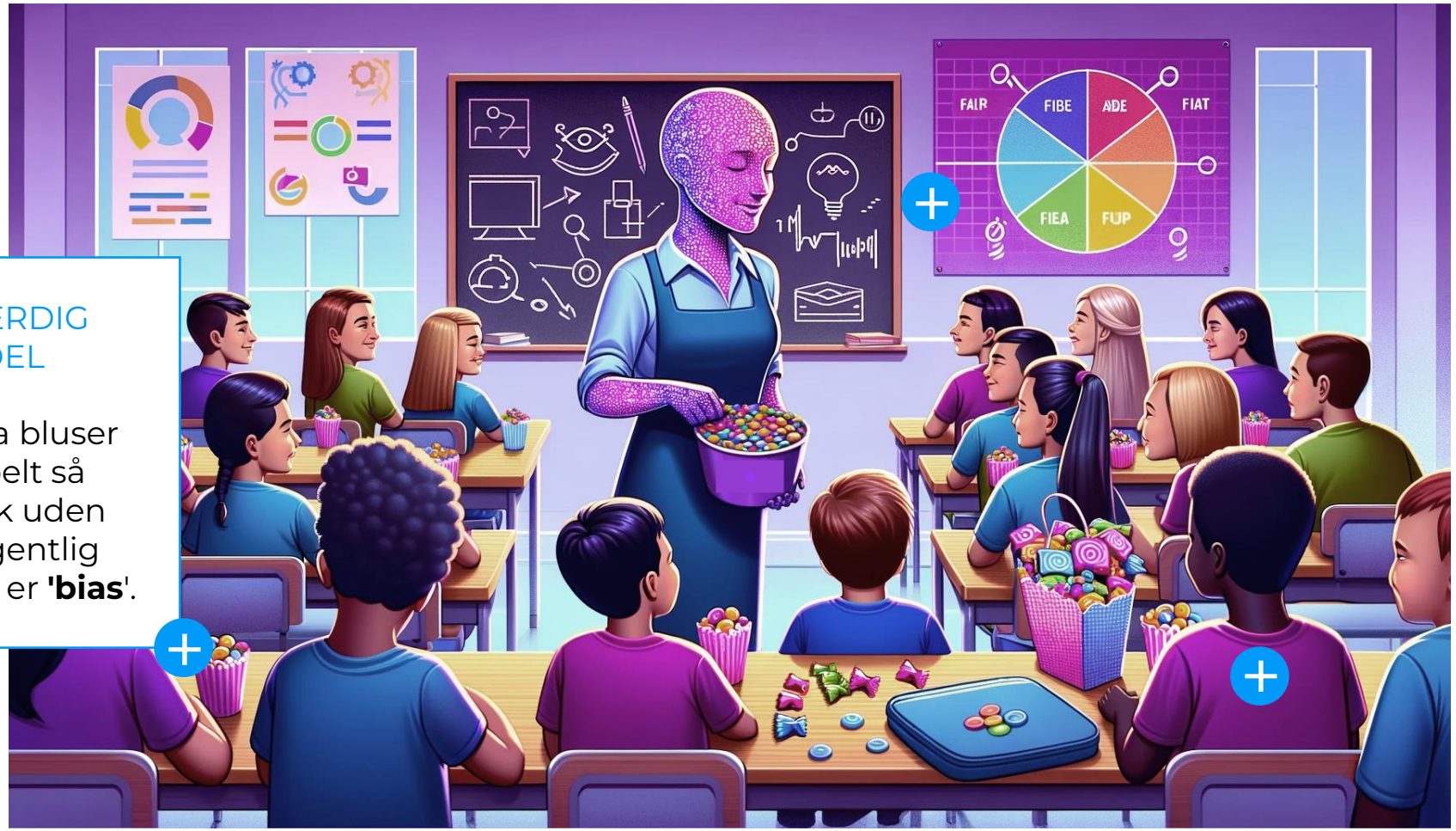
Forestil dig et klasseværelse, hvor dine chancer for at få ekstra slik afhænger af, hvad du har på!

Introduktion til data-fairness og bias

Fairness og bias – slikanalogien

URETFÆRDIG FORDEL

Elever i lilla bluser får dobbelt så meget slik uden nogen egentlig grund - det er **'bias**'.



BILLEDKILDE | Genereret af DALL-E

Introduktion til data-fairness og bias

Fairness og bias – slikanalogien

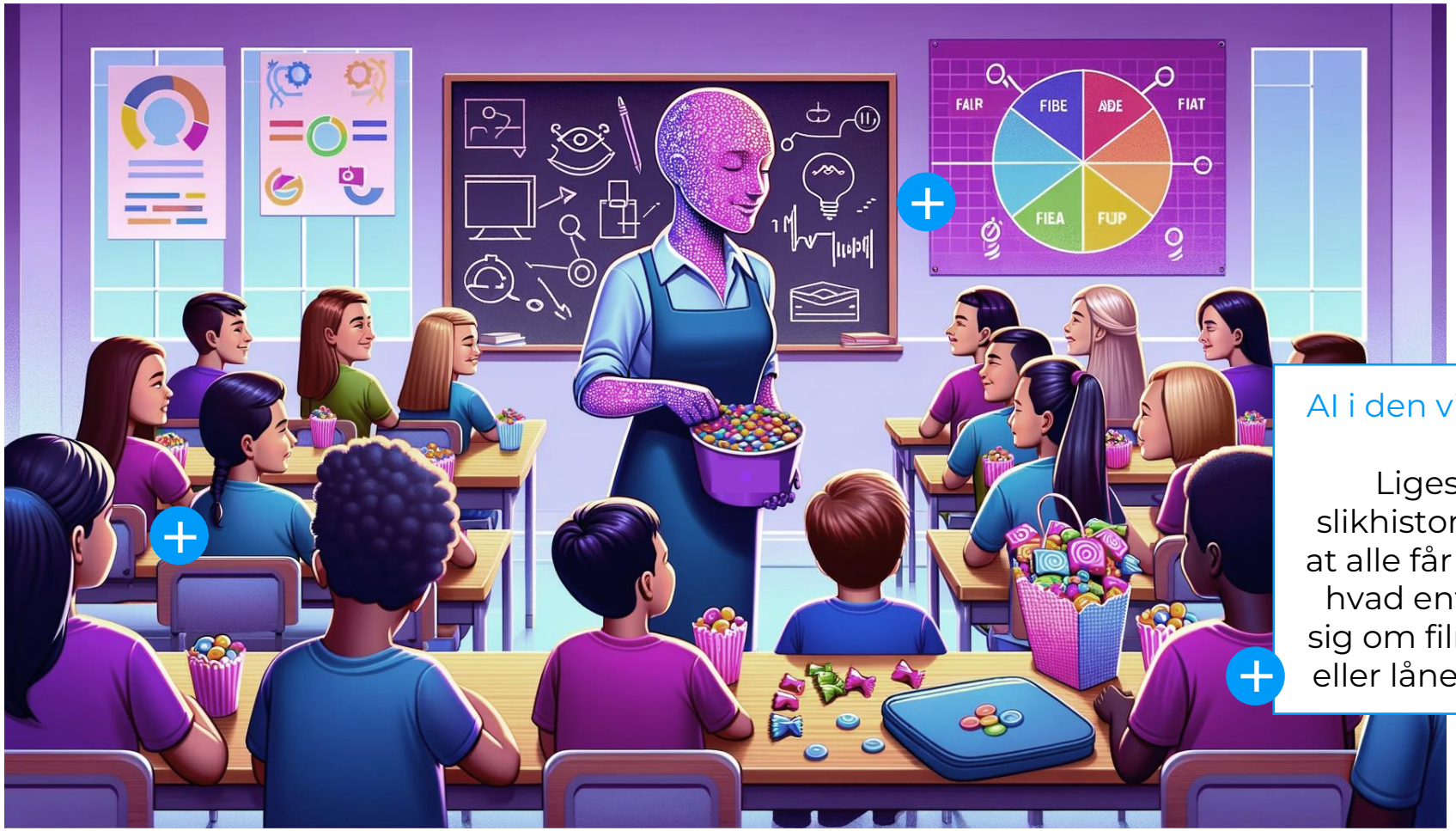
LIGE MULIGHEDER

Hvis løsningen af et puslespil afgør, hvem der får ekstra slik, er alle på lige fod - det er **'fairness**'.

BILLEDKILDE | Genereret af DALL-E

Introduktion til data-fairness og bias

Fairness og bias – slikanalogien



AI i den virkelige verden

Ligesom i vores slikhistorie bør AI sikre, at alle får en fair chance, hvad enten det drejer sig om filmanbefalinger eller lånegodkendelser.

BILLEDKILDE | Genereret af DALL-E

Introduktion til data-fairness og bias

Retfærdige data

Dataretfærdighed handler om at sikre, at alle individer behandles retfærdigt af AI-systemer, uden at bias påvirker systemernes resultater. Det involverer de processer og praksisser, der indgår i indsamling, behandling og brug af data for at forhindre diskrimination.

Retfærdige data er et afgørende aspekt af at designe og implementere AI-systemer, der har en positiv indvirkning på samfundet. Det sikrer, at AI-beslutninger er retfærdige og upartiske, fremmer lighed og forhindrer diskrimination.

Fair datapraksis fører til pålidelige AI-applikationer, der kan accepteres bredt og være til gavn for forskellige samfundsgrupper.

Den sigter mod AI-systemer, der tjener alle ligeligt, uanset baggrund eller demografiske karakteristika.

Introduktion til data-fairness og bias

- Hvad er bias? +
- Hvordan opstår det? +
- Hvad er konsekvenserne af bias? +

Introduktion til data-fairness og bias

Hvad er bias?

Bias i AI refererer til systematiske fejl, der uretfærdigt favoriserer en gruppe frem for andre. Det kan skyldes skæve data, fejlbehæftede algoritmer eller træningsprocesser, der er præget af fordomme.

Hvordan opstår det? +

Hvad er konsekvenserne af bias? +

Introduktion til data-fairness og bias

Hvad er bias? +

Hvordan opstår det? -

Data bias: Når data, der indføres i AI-systemer, afspejler historiske uligheder eller ikke-repræsentative samples fra forskellige befolkningsgrupper.
Algoritme-bias: Når de regler eller modeller, der bruges til at træffe beslutninger, forstærker eksisterende fordomme eller overser minoriteter.
Feedback-loops: når forudindtagede AI-beslutninger påvirker fremtidige data og fortsætter den forudindtagede cyklus.

Hvad er konsekvenserne af bias? +

Introduktion til data-fairness og bias

Hvad er bias? +

Hvordan opstår det? +

Hvad er konsekvenserne af bias? -

Sociale konsekvenser: Kan føre til opretholdelse af stereotyper, forstærke sociale uligheder og underminere tilliden til vigtige institutioner.

Konsekvenser for individet: Folk kan blive udsat for uretfærdig behandling baseret på fejlbehæftede AI-beslutninger inden for vigtige områder som ansættelse, retshåndhævelse og lånegodkendelse.

Introduktion til data-fairness og bias

Scenarier fra den virkelige verden: AI-bias

Bias i rekrutteringsværktøjet

Scenarie

I 2018 blev det kom det frem, at Amazon havde skrottet et AI-rekrutteringsværktøj, der udviste partiskhed over for kvinder. AI-systemet blev trænet på CV-data indsamlet over en 10-årig periode, overvejende fra mænd, hvilket afspejler den mandlige dominans i teknologibranchen. Derfor lærte AI'en sig selv, at mandlige kandidater var at foretrække, og nedprioriterede CV'er, der indeholdt ord som "kvinde", som f.eks. i ”leder af skakklub for kvinder”.

Konsekvens

Dette eksempel fremhæver, hvordan AI kan fastholde historiske fordomme, hvis træningsdataene er forudindtagede eller biased. Det viser også, hvordan AI kan påvirke beskæftigelsesmulighederne, og hvor vigtigt det er at sikre, at AI-rekrutteringsværktøjer fremmer mangfoldighed og retfærdighed.

Introduktion til data-fairness og bias

Scenarier fra den virkelige verden: AI-bias

Bias i sundhedsalgoritmer

Scenarie

En undersøgelse fra 2019 offentliggjort i ”Science” afslørede, at en algoritme, der blev brugt i mange amerikanske sundhedssystemer til at fordele sundhedsressourcer, var partisk over for sorte patienter. Algoritmen forudsagde personers behov for sundhedsydelser baseret på sundhedsomkostninger og antog utilsigtet, at lavere sundhedsomkostninger betød lavere behov for ydelser. Men historisk set har sorte patienter haft færre omkostninger på grund af forskellige barrierer for at få adgang til behandling, ikke fordi de var sundere. Som følge heraf favoriserede algoritmen sundere hvide patienter frem for mere syge sorte patienter, når den skulle identificere kandidater til yderligere pleje.

Konsekvens

Denne skævhed i algoritmen resulterede i, at sorte patienter fik betydeligt mindre adgang til pleje, der er designet til at hjælpe med at håndtere komplekse sundhedstilstande, selvom deres behov for sådanne interventioner var højere. Dette scenarie understreger det vigtige behov for retfærdighed i AI-applikationer i sundhedssektoren, hvor forudindtagede beslutninger kan have direkte indflydelse på liv og trivsel.

TYPER AF BIAS I AI-SYSTEMER

BILLEDKILDE | Genereret af DALL-E

Typer af bias i AI-systemer

Bias i udvælgelsen

Klik på kort for at vende det

Selektionsbias

Klik på kort for at vende det

Målingsbias

Klik på kort for at vende det

Algoritmisk bias

Klik på kort for at vende det

Bekræftelsesbias

Klik på kort for at vende det

Bias i rapporteringen

Klik på kort for at vende det

Typer af bias i AI-systemer

Udvælgelsesbias opstår, når det data-udsnit, der bruges til analyse, ikke er repræsentativt for den befolkning, det skal repræsentere. Det fører til skæve eller unøjagtige konklusioner.

Selektionsbias

Klik på kort for at vende det

Målingsbias

Klik på kort for at vende det

Algoritmisk bias

Klik på kort for at vende det

Bekræftelsesbias

Klik på kort for at vende det

Bias i rapporteringen

Klik på kort for at vende det

Typer af bias i AI-systemer

Udvælgelsesbias opstår, når det data-udsnit, der bruges til analyse, ikke er repræsentativt for den befolkning, det skal repræsentere. Det fører til skæve eller unøjagtige konklusioner.

Selektionsbias er en delmængde udvælgelsesbias og opstår, når den metode, der bruges til at indsamle data, favoriserer visse typer individer eller grupper frem for andre.

Målingsbias

Klik på kort for at vende det

Algoritmisk bias

Klik på kort for at vende det

Bekræftelsesbias

Klik på kort for at vende det

Bias i rapporteringen

Klik på kort for at vende det

Typer af bias i AI-systemer

Udvælgelsesbias opstår, når det data-udsnit, der bruges til analyse, ikke er repræsentativt for den befolkning, det skal repræsentere. Det fører til skæve eller unøjagtige konklusioner.

Selektionsbias er en delmængde udvælgelsesbias og opstår, når den metode, der bruges til at indsamle data, favoriserer visse typer individer eller grupper frem for andre.

Målingsbias opstår på grund af fejl eller uoverensstemmelser i måleprocessen, hvilket fører til unøjagtige eller vildledende resultater. Det kan skyldes defekte instrumenter, tvetydige spørgsmål eller observatørens bias.

Algoritmisk bias

Klik på kort for at vende det

Bekræftelsesbias

Klik på kort for at vende det

Bias i rapporteringen

Klik på kort for at vende det

Typer af bias i AI-systemer

Udvælgelsesbias opstår, når det data-udsnit, der bruges til analyse, ikke er repræsentativt for den befolkning, det skal repræsentere. Det fører til skæve eller unøjagtige konklusioner.

Selektionsbias er en delmængde udvælgelsesbias og opstår, når den metode, der bruges til at indsamle data, favoriserer visse typer individer eller grupper frem for andre.

Målingsbias opstår på grund af fejl eller uoverensstemmelser i måleprocessen, hvilket fører til unøjagtige eller vildledende resultater. Det kan skyldes defekte instrumenter, tvetydige spørgsmål eller observatørens bias.

Algoritmisk bias opstår i maskinlæringsalgoritmer, når de systematisk diskriminerer visse grupper eller individer baseret på race, køn eller andre beskyttede karakteristika, der findes i træningsdataene.

Bekræftelsesbias

Klik på kort for at vende det

Bias i rapporteringen

Klik på kort for at vende det

Typer af bias i AI-systemer

Udvælgelsesbias opstår, når det data-udsnit, der bruges til analyse, ikke er repræsentativt for den befolkning, det skal repræsentere. Det fører til skæve eller unøjagtige konklusioner.

Selektionsbias er en delmængde udvælgelsesbias og opstår, når den metode, der bruges til at indsamle data, favoriserer visse typer individer eller grupper frem for andre.

Målingsbias opstår på grund af fejl eller uoverensstemmelser i måleprocessen, hvilket fører til unøjagtige eller vildledende resultater. Det kan skyldes defekte instrumenter, tvetydige spørgsmål eller observatørens bias.

Algoritmisk bias opstår i maskinlæringsalgoritmer, når de systematisk diskriminerer visse grupper eller individer baseret på race, køn eller andre beskyttede karakteristika, der findes i træningsdataene.

Bekræftelsesbias er, når forskere eller analytikere selektivt bruger eller fortolker data, der bekræfter deres forudfattede meninger eller hypoteser, mens de ignorerer eller afviser modstridende beviser eller fakta.

Bias i rapporteringen

Klik på kort for at vende det

Typer af bias i AI-systemer

Udvælgelsesbias opstår, når det data-udsnit, der bruges til analyse, ikke er repræsentativt for den befolkning, det skal repræsentere. Det fører til skæve eller unøjagtige konklusioner.

Selektionsbias er en delmængde udvælgelsesbias og opstår, når den metode, der bruges til at indsamle data, favoriserer visse typer individer eller grupper frem for andre.

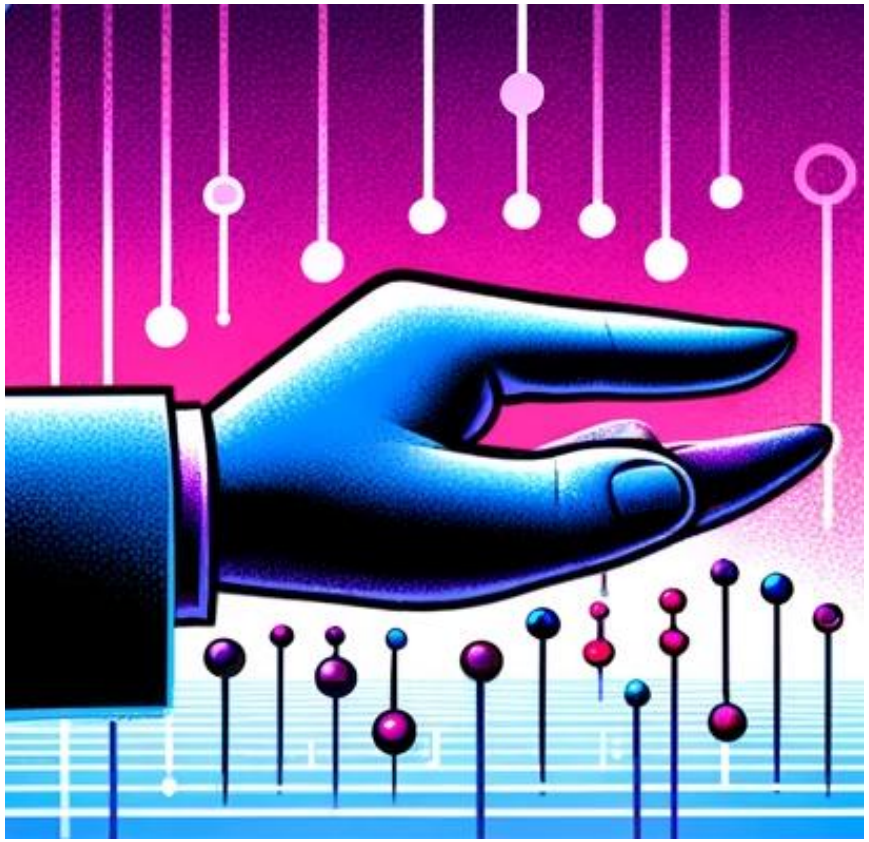
Målingsbias opstår på grund af fejl eller uoverensstemmelser i måleprocessen, hvilket fører til unøjagtige eller vildledende resultater. Det kan skyldes defekte instrumenter, tvetydige spørgsmål eller observatørens bias.

Algoritmisk bias opstår i maskinlæringsalgoritmer, når de systematisk diskriminerer visse grupper eller individer baseret på race, køn eller andre beskyttede karakteristika, der findes i træningsdataene.

Bekræftelsesbias er, når forskere eller analytikere selektivt bruger eller fortolker data, der bekræfter deres forudfattede meninger eller hypoteser, mens de ignorerer eller afviser modstridende beviser eller fakta.

Bias i rapporteringen refererer til selektiv offentliggørelse eller afrapportering af forskningsresultater, der er statistisk signifikante eller gunstige, mens ikke-signifikante eller ugunstige resultater undertrykkes eller ignoreres.

Typer af bias i AI-systemer
Eksempler



BILLEDKILDE | Genereret af DALL-E

Selektionsbias | Ansigtsgenkendelsesteknologi

Ansigtsgenkendelsesteknologier er ofte blevet trænet på datasæt, der hovedsageligt består af kaukasiske ansigter. Denne mangel på diversitet i træningsdatasættene har ført til højere fejlprocenter ved genkendelse af personer med anden etnisk baggrund. For eksempel viste en undersøgelse fra MIT Media Lab, at der var betydelige forskelle i nøjagtigheden af kønsgenkendelsesteknologier, især fejlidentificering af køn blandt kvinder med mørkere hudfarve.

Typer af bias i AI-systemer
Eksempler

BILLEDKILDE | Genereret af DALL-E

Selektionsbias | Fejl i politiske meningsmålinger

Ved det amerikanske præsidentvalg i 2016 undervurderede flere meningsmålinger støtten til Donald Trump. Et centralt problem var bias i stikprøverne, hvor den demografiske vægtning i stikprøverne ikke repræsenterede vælgerbasen præcist. Meningsmålingsinstitutterne havde taget for mange stikprøver i byerne og for få i landområderne, hvilket førte til unøjagtige forudsigelser, der ikke ramte plet i forhold til det faktiske valgresultat.

Typer af bias i AI-systemer
Eksempler

BILLEDKILDE | Genereret af DALL-E

Målingsbias | Pulsoximetre og hudfarve

Nylige undersøgelser, herunder dem, der blev forstærket af COVID-19-pandemien, har vist, at pulsoximetre (puls- og iltmålere) kan udvise målingsbias ved at overestimere iltniveauet i blodet hos mennesker med mørkere hud. Denne bias opstår, fordi enhedernes sensorer er mindre effektive til at trænge igennem mørkere hud, hvilket kan føre til farlige fejldiagnoser eller at behandlingen af farvede patienter forsinkes.

Typer af bias i AI-systemer
Eksempler

BILLEDKILDE | Genereret af DALL-E

Algoritmisk bias | COMPAS i strafudmåling

COMPAS-softwaren (Correctional Offender Management Profiling for Alternative Sanctions), der bruges i det amerikanske strafferetssystem til at vurdere sandsynligheden for, at en tiltalt begår ny kriminalitet, viste sig at have racemæssige bias. Analysen afslørede, at softwaren var næsten dobbelt så tilbøjelig til fejlagtigt at forudsige fremtidig kriminalitet hos sorte tiltalte sammenlignet med hvide tiltalte, hvilket påvirkede beslutninger om strafudmåling og kaution baseret på skæve algoritmiske anbefalinger.

Typer af bias i AI-systemer
Eksempler

BILLEDKILDE | Genereret af DALL-E

Bekræftelsesbias - kreditvurderingsmodeller

Finansielle institutioner, der bruger AI-drevne kreditvurderingsmodeller, kan utilsigtet forstærke bekræftelsesbias, hvis modellerne er trænet på historiske udlånsdata, der afspejler tidligere fordomme. Hvis tidligere menneskelige långivere f.eks. havde fordomme mod visse demografiske grupper, kan AI-systemet gentage disse fordomme ved at favorisere personer fra historisk favoriserede grupper og antage, at de er mindre risikable baseret på tidligere tilbagebetalinger af lån.

Typer af bias i AI-systemer
Eksempler

BILLEDKILDE | Genereret af DALL-E

Bias i rapporteringen - Lægemiddelbivirkninger i onlinefora

Medicinalvirksomheder og sundhedsforskere, der bruger AI til at overvåge onlinefora og sociale medier for rapporter om lægemiddelbivirkninger, kan støde på rapporteringsbias. Patienter, der oplever alvorlige bivirkninger, er mere tilbøjelige til at rapportere deres oplevelser online sammenlignet med dem, der oplever milde eller ingen bivirkninger. Det kan få AI-systemer til at overvurdere sværhedsgraden eller hyppigheden af bivirkninger, hvilket potentielt kan påvirke lægemidlers sikkerhedsprofiler på en unøjagtig måde.

METODER TIL AT IDENTIFICERE OG MÅLE BIAS I DATA

BILLEDKILDE | Genereret af DALL-E

Metoder til at identificere og måle bias i data

- Revision af data +
- Analyse af forskelle (disparitetsanalyse) +
- Algoritmisk fairness-test +
- Sensivitetsanalyse +

Metoder til at identificere og måle bias i data

Revision af data -

Datarevision indebærer systematisk undersøgelse af datasæt for at identificere eventuelle forskelle, ubalancer eller afvigelser, der kan indikere bias.

Denne proces omfatter kontrol af repræsentativ lighed mellem forskellige grupper i dataene. Den sikrer, at dataindsamlingsprocessen ikke systematisk udelukker visse befolkningsgrupper eller forstærker eksisterende stereotyper.

Analyse af forskelle (disparitetsanalyse) +

Algoritmisk fairness-test +

Sensivitetsanalyse +

Metoder til at identificere og måle bias i data

Revision af data +

Analyse af forskelle (disparitetsanalyse) -

Disparitetsanalyse refererer til at sammenligne resultaterne eller forudsigelserne af AI-modeller på tværs af forskellige demografiske eller beskyttede grupper for at identificere eventuelle ulige virkninger.

Det indebærer beregning af statistiske mål som f.eks. forskellen i fejlprocenter, acceptprocenter eller andre resultatmålinger mellem grupper. Det hjælper med at kvantificere og visualisere forskelle, der kan indikere bias i AI-beslutningsprocesser.

Algoritmisk fairness-test +

Sensivitetsanalyse +

Metoder til at identificere og måle bias i data

Revision af data +

Analyse af forskelle (disparitetsanalyse) +

Algoritmisk fairness-test -

Algoritmisk retfærdighedstestning indebærer implementering af test eller målinger, der vurderer retfærdigheden af beslutninger truffet af AI-systemer.

Det kan omfatte brug af retfærdighedsmålinger som f.eks. lige muligheder, demografisk paritet eller individuelle retfærdighedskriterier til at evaluere algoritmernes performance. Disse tests hjælper med at afgøre, om AI-systemer træffer retfærdige og upartiske beslutninger på tværs af forskellige demografiske grupper.

Sensivitetsanalyse +

Metoder til at identificere og måle bias i data

Revision af data +

Analyse af forskelle (disparitetsanalyse) +

Algoritmisk fairness-test +

Sensivitetsanalyse -

Sensivitetsanalyse indebærer, at man tester robustheden af AI-beslutninger ved at ændre lidt på datainput og observere, hvordan resultaterne varierer for forskellige grupper.

Ved at 'forstyrre' datainputtene vurderer Sensivitetsanalysen, hvor følsomme AI-modeller er over for ændringer i inputdataene. Det hjælper med at identificere potentielle kilder til bias og vurdere pålideligheden og konsistensen af AI-forudsigelser på tværs af forskellige datasæt.

Metoder til at identificere og måle bias i data

Interaktiv workshop om at identificere og måle bias i data

Bias i rekrutteringsværktøj

Mål: Denne workshop vil sætte deltagerne i stand til at anvende datarevision, disparitetsanalyse og sensivitetsanalyse ved hjælp af et virkeligt datasæt for at identificere og forstå potentielle bias.

Værktøj og materialer

Datasæt: Giv deltagerne et enkelt, offentligt tilgængeligt datasæt med forskellige demografiske egenskaber (f.eks. datasættet Adult Income fra UCI Machine Learning Repository, som indeholder karakteristika som alder, uddannelse, race, køn og indkomst).

Aktivitet: Deltagerne indlæser datasættet og udfører grundlæggende dataudforskning (tjekker for manglende værdier og forstår fordelingen af nøgleattributter).

Mål: At gøre deltagerne fortrolige med datasættet og forberede data til analyse.

Datarevision: Gennemfør en datarevision ved at analysere repræsentationen af forskellige grupper i datasættet. Beregn andelen af hver demografisk gruppe (f.eks. efter køn, race). Identificer eventuelle underrepræsenterede grupper.

Disparitetsanalyse: Deltagerne anvender disparitetsanalyse til at vurdere, hvor godt en simpel model (f.eks. logistisk regression til forudsigelse af indkomstniveau) forudsiger forskellige grupper. Identificer eventuelle betydelige forskelle i modellens performance på tværs af grupper, og opstil hypoteser om potentielle årsager.

Sensivitetsanalyse: Udfør en sensivitetssanalyse ved at ændre lidt på datapunkterne og observer ændringer i modellens forudsigelser. Ændr attributter relateret til følsomme demografiske variabler (f.eks. ændr alder eller uddannelsesniveau en smule), og bemærk, hvordan ændringerne påvirker modellens forudsigelser.

STRATEGIER TIL AT HÅNDTERE OG AFBØDE BIAS

BILLEDKILDE | Genereret af DALL-E

Strategier til at håndtere og afbøde bias

Mangfoldig repræsentation

Klik på kort for at vende det

Bias-bevidste algoritmer

Klik på kort for at vende det

Regelmæssig overvågning og evaluering

Klik på kort for at vende det

Gennemsigtighed og forklarlighed

Klik på kort for at vende det

Strategier til at håndtere og afbøde bias

Mangfoldig repræsentation

Forskellige datapunkter fra forskellige demografiske grupper sikrer, at AI-systemer trænes på et repræsentativt datasæt, hvilket reducerer sandsynligheden for forudindtagede resultater. Mangfoldighed giver forskellige perspektiver og indsigter, der hjælper med at identificere og afbøde bias under udviklingsprocessen.

Bias-bevidste algoritmer

Klik på kort for at vende det

Regelmæssig overvågning og evaluering

Klik på kort for at vende det

Gennemsigtighed og forklarlighed

Klik på kort for at vende det

Strategier til at håndtere og afbøde bias

Mangfoldig repræsentation

Forskellige datapunkter fra forskellige demografiske grupper sikrer, at AI-systemer trænes på et repræsentativt datasæt, hvilket reducerer sandsynligheden for forudindtagede resultater. Mangfoldighed giver forskellige perspektiver og indsigter, der hjælper med at identificere og afbøde bias under udviklingsprocessen.

Bias-bevidste algoritmer

Design af algoritmer med indbyggede mekanismer til at opdage og afbøde bias. Det kan indebære at indarbejde fairness-begrænsninger i algoritmens optimeringsproces eller at bruge teknikker som *adversarial training* til at minimere bias i modelforudsigelser.

Regelmæssig overvågning og evaluering

Klik på kort for at vende det

Gennemsigtighed og forklarlighed

Klik på kort for at vende det

Strategier til at håndtere og afbøde bias

Mangfoldig repræsentation

Forskellige datapunkter fra forskellige demografiske grupper sikrer, at AI-systemer trænes på et repræsentativt datasæt, hvilket reducerer sandsynligheden for forudindtagede resultater. Mangfoldighed giver forskellige perspektiver og indsigter, der hjælper med at identificere og afbøde bias under udviklingsprocessen.

Bias-bevidste algoritmer

Design af algoritmer med indbyggede mekanismer til at opdage og afbøde bias. Det kan indebære at indarbejde fairness-begrænsninger i algoritmens optimeringsproces eller at bruge teknikker som *adversarial training* til at minimere bias i modelforudsigelser.

Regelmæssig overvågning og evaluering

Løbende overvågning af AI-systemer i den virkelige verden for at identificere bias, når de opstår. Regelmæssig evaluering omfatter analyse af AI-output for forskelle på tværs af forskellige demografiske grupper og opdatering af algoritmer eller datasæt for at imødegå eventuelle identificerede skævheder med det samme.

Gennemsigtighed og forklarlighed

Klik på kort for at vende det

Strategier til at håndtere og afbøde bias

Mangfoldig repræsentation

Forskellige datapunkter fra forskellige demografiske grupper sikrer, at AI-systemer trænes på et repræsentativt datasæt, hvilket reducerer sandsynligheden for forudindtagede resultater. Mangfoldighed giver forskellige perspektiver og indsigter, der hjælper med at identificere og afbøde bias under udviklingsprocessen.

Bias-bevidste algoritmer

Design af algoritmer med indbyggede mekanismer til at opdage og afbøde bias. Det kan indebære, at der indarbejdes fairness-begrænsninger i algoritmens optimeringsproces, eller at der bruges teknikker som *adversarial training* til at minimere bias i modelforudsigelser.

Regelmæssig overvågning og evaluering

Løbende overvågning af AI-systemer i den virkelige verden for at identificere bias, når de opstår. Regelmæssig evaluering omfatter analyse af AI-output for forskelle på tværs af forskellige demografiske grupper og opdatering af algoritmer eller datasæt for at imødegå eventuelle identificerede skævheder med det samme.

Gennemsigtighed og forklarlighed

Gennemsigtighed og forklaringer på AI-beslutninger giver interessenter mulighed for at forstå, hvordan beslutninger træffes, og identificere eventuelle bias i systemet. AI-teknikker, der kan forklares, f.eks. ved at give en score for vigtighed af funktioner eller en begrundelse for beslutninger, gør det muligt for brugerne at fortolke og vurdere, om AI-resultaterne er retfærdige.

FORSTÅELSE AF FAIRNESS-METRIKKER

BILLEDKILDE | Genereret af DALL-E

Forståelse af fairness-metrikker

Statistisk paritet	+
Udlignede odds (*equalized odds*)	+
Betinget demografisk forskel	+
Retfærdighed gennem bevidsthed	+
Individuel retfærdighed	+
Kontrafaktisk retfærdighed	+

Forståelse af fairness-metrikker

Statistisk paritet

Statistisk paritet er også kendt som demografisk paritet, og denne metrik vurderer, om andelen af positive resultater (f.eks. lånegodkendelser eller jobtilbud) er den samme på tværs af forskellige demografiske grupper.

Udlignede odds (*equalized odds*) +

Betinget demografisk forskel +

Retfærdighed gennem bevidsthed +

Individuel retfærdighed +

Kontrafaktisk retfærdighed +

Forståelse af fairness-metrikker

Statistisk paritet +

Udlignede odds (*equalized odds*)

Equalized odds undersøger, om den sande positive rate (sensitivitet) og den sande negative rate (specificitet) er ens på tværs af forskellige demografiske grupper. Det sikrer, at modellens prædiktive performance er konsistent på tværs af alle grupper.

Betinget demografisk forskel +

Retfærdighed gennem bevidsthed +

Individuel retfærdighed +

Kontrafaktisk retfærdighed +

Forståelse af fairness-metrikker

Statistisk paritet +

Udlignede odds (*equalized odds*) +

Betinget demografisk forskel:

Betinget demografisk forskel evaluerer uligheden i forudsigelige resultater betinget af en beskyttet egenskab. Den måler forskellen i forudsigelsesresultater mellem forskellige demografiske grupper, idet der tages højde for andre relevante faktorer.

Retfærdighed gennem bevidsthed +

Individuel retfærdighed +

Kontrafaktisk retfærdighed +

Forståelse af fairness-metrikker

Statistisk paritet +

Udlignede odds (*equalized odds*) +

Betinget demografisk forskel +

Retfærdighed gennem bevidsthed

Retfærdighed gennem bevidsthed inddrager bevidsthed om følsomme egenskaber (f.eks. race eller køn) i beslutningsprocessen. Det sikrer, at beslutninger truffet af AI-systemer er retfærdige og upartiske, idet der tages højde for den historiske kontekst og de samfundsmæssige konsekvenser.

Individuel retfærdighed +

Kontrafaktisk retfærdighed +

Forståelse af fairness-metrikker

Statistisk paritet +

Udlignede odds (*equalized odds*) +

Betinget demografisk forskel +

Retfærdighed gennem bevidsthed +

Individuel retfærdighed

Individuel retfærdighed fokuserer på at sikre, at lignende personer får samme behandling eller resultater fra AI-systemet, uanset deres demografiske karakteristika. Det sigter mod at minimere ulige behandling baseret på irrelevante egenskaber.

Kontrafaktisk retfærdighed +

Forståelse af fairness-metrikker

Statistisk paritet +

Udlignede odds (*equalized odds*) +

Betinget demografisk forskel +

Retfærdighed gennem bevidsthed +

Individuel retfærdighed +

Kontrafaktisk retfærdighed

Kontrafaktisk retfærdighed vurderer, om en ændring af en persons følsomme egenskab (f.eks. race eller køn), mens andre egenskaber holdes konstante, vil resultere i en ændring af beslutningsresultatet. Det sikrer, at enkeltpersoner behandles retfærdigt uanset deres beskyttede egenskaber.

Forståelse af fairness-metrikker

Interaktiv quiz

Oprettelse af en interaktiv quiz på Mentimeter

Afstemningsspørgsmål

Før vi går i gang, hvad er så vigtige faktorer at overveje, når man skal afgøre, om et AI-system er retfærdigt?

Valgmuligheder

a) Retfærdige resultater; b) Demografisk repræsentation; c) Ligebehandling; d) Gennemsigtighed.

Forståelse af fairness-metrikker

Interaktive klasseøvelser

Emne: Statistisk paritet

Beregn statistisk paritet

- På baggrund af dataene skal du beregne andelen af godkendte lån for hver race. Er der paritet?
- Giv deltagerne tallene (eller procenterne) for godkendelser og afvisninger for hver race, og bed dem om at beregne, om satserne er nogenlunde ens.
- Multiple choice-svar: "Satserne er ens; paritet er opnået." og "Satserne er ikke ens; der er en forskel."

Forståelse af fairness-metrikker

Interaktive klasseøvelser

Emne: Udlignede odds

Vurder udlignede odds

- Hvis vi antager, at den sande positive rate (korrekt godkendt lån til dem, der burde være godkendt) for hvide er 80 % og for sorte 60 %, er der så udlignede odds?
- Multiple choice-svar: "Ja, oddsene er lige store." eller "Nej, oddsene er ikke lige store."

RETFÆRDIGHED I DATAINDSAMLING OG FORBEHANDLING

BILLEDKILDE | Genereret af DALL-E

Retfærdighed i dataindsamling og forbehandling

Forskellige datakilder

Klik på kort for at vende det

Inkluderende dataindsamling

Klik på kort for at vende det

Rensning af data

Klik på kort for at vende det

Registrering af bias

Klik på kort for at vende det

Retfærdig funktionsudvikling

Klik på kort for at vende det

Gennemsigtighed og dokumentation

Klik på kort for at vende det

Regelmæssig overvågning

Klik på kort for at vende det

Retfærdighed i dataindsamling og forbehandling

Forskellige datakilder

Indsaml data fra forskellige kilder for at sikre en omfattende repræsentation af befolkningen. Det er med til at forhindre underrepræsentation eller skævvridning af specifikke demografiske grupper.

Inkluderende dataindsamling

Klik på kort for at vende det

for at afhjælpe ubalancer.

Rensning af data

Klik på kort for at vende det

standardisering af dataformater.

Registrering af bias

Klik på kort for at vende det

grupper.

Retfærdig funktionsudvikling

Klik på kort for at vende det

Gennemsigtighed og dokumentation

Klik på kort for at vende det

eller feature engineering, som kan påvirke retfærdigheden.

Regelmæssig overvågning

Klik på kort for at vende det

sikre retfærdighed over tid.

Retfærdighed i dataindsamling og forbehandling

Forskellige datakilder

Indsaml data fra forskellige kilder for at sikre en omfattende repræsentation af befolkningen. Det er med til at forhindre underrepræsentation eller skævvridning af specifikke demografiske grupper.

Inkluderende dataindsamling

Brug inkluderende dataindsamlingsteknikker for at sikre, at alle dele af befolkningen er tilstrækkeligt repræsenteret i datasættet. Det kan indebære stratificeret sampling eller oversampling af minoritetsgrupper for at afhjælpe ubalancer.

Rensning af data

Klik på kort for at vende det

standardisering af dataformater.

Registrering af bias

Klik på kort for at vende det

grupper.

Retfærdig funktionsudvikling

Klik på kort for at vende det

Gennemsigtighed og dokumentation

Klik på kort for at vende det

eller feature engineering, som kan påvirke retfærdigheden.

Regelmæssig overvågning

Klik på kort for at vende det

sikre retfærdighed over tid.

Retfærdighed i dataindsamling og forbehandling

Forskellige datakilder

Indsaml data fra forskellige kilder for at sikre en omfattende repræsentation af befolkningen. Det er med til at forhindre underrepræsentation eller skævvridning af specifikke demografiske grupper.

Inkluderende dataindsamling

Brug inkluderende dataindsamlingsteknikker for at sikre, at alle dele af befolkningen er tilstrækkeligt repræsenteret i datasættet. Det kan indebære stratificeret sampling eller oversampling af minoritetsgrupper for at afhjælpe ubalancer.

Rensning af data

Rens og forarbejd dataene omhyggeligt for at fjerne eventuelle skævheder eller uoverensstemmelser. Dette omfatter identifikation og korrektion af fejl, håndtering af manglende værdier og standardisering af dataformater.

Registrering af bias

Klik på kort for at vende det

grupper.

Retfærdig funktionsudvikling

Klik på kort for at vende det

Gennemsigtighed og dokumentation

Klik på kort for at vende det

eller feature engineering, som kan påvirke retfærdigheden.

Regelmæssig overvågning

Klik på kort for at vende det

sikre retfærdighed over tid.

Retfærdighed i dataindsamling og forbehandling

Forskellige datakilder

Indsaml data fra forskellige kilder for at sikre en omfattende repræsentation af befolkningen. Det er med til at forhindre underrepræsentation eller skævvridning af specifikke demografiske grupper.

Inkluderende dataindsamling

Brug inkluderende dataindsamlingsteknikker for at sikre, at alle dele af befolkningen er tilstrækkeligt repræsenteret i datasættet. Det kan indebære stratificeret sampling eller oversampling af minoritetsgrupper for at afhjælpe ubalancer.

Rensning af data

Rens og forarbejd dataene omhyggeligt for at fjerne eventuelle skævheder eller uoverensstemmelser. Dette omfatter identifikation og korrektion af fejl, håndtering af manglende værdier og standardisering af dataformater.

Registrering af bias

Implementer teknikker til at opdage og afbøde bias i data. Det kan indebære, at man gennemfører bias-audits eller bruger automatiserede værktøjer til at identificere forskelle på tværs af forskellige demografiske grupper.

Retfærdig funktionsudvikling

Klik på kort for at vende det

Gennemsigtighed og dokumentation

Klik på kort for at vende det

eller feature engineering, som kan påvirke retfærdigheden.

Regelmæssig overvågning

Klik på kort for at vende det

sikre retfærdighed over tid.

Retfærdighed i dataindsamling og forbehandling

Forskellige datakilder

Indsaml data fra forskellige kilder for at sikre en omfattende repræsentation af befolkningen. Det er med til at forhindre underrepræsentation eller skævvridning af specifikke demografiske grupper.

Inkluderende dataindsamling

Brug inkluderende dataindsamlingsteknikker for at sikre, at alle dele af befolkningen er tilstrækkeligt repræsenteret i datasættet. Det kan indebære stratificeret sampling eller oversampling af minoritetsgrupper for at afhjælpe ubalancer.

Rensning af data

Rens og forarbejd dataene omhyggeligt for at fjerne eventuelle skævheder eller uoverensstemmelser. Dette omfatter identifikation og korrektion af fejl, håndtering af manglende værdier og standardisering af dataformater.

Registrering af bias

Implementer teknikker til at opdage og afbøde bias i data. Det kan indebære, at man gennemfører bias-audits eller bruger automatiserede værktøjer til at identificere forskelle på tværs af forskellige demografiske grupper.

Retfærdig funktionsudvikling

Overvej konsekvenserne af valg af funktioner og teknik for retfærdighed. Undgå at bruge funktioner, der kan fastholde stereotyper eller diskrimination, og stræb efter funktioner, der er relevante og ikke-diskriminerende.

Gennemsigtighed og dokumentation

Klik på kort for at vende det

eller feature engineering, som kan påvirke retfærdigheden.

Regelmæssig overvågning

Klik på kort for at vende det

sikre retfærdighed over tid.

Retfærdighed i dataindsamling og forbehandling

Forskellige datakilder

Indsaml data fra forskellige kilder for at sikre en omfattende repræsentation af befolkningen. Det er med til at forhindre underrepræsentation eller skævvridning af specifikke demografiske grupper.

Inkluderende dataindsamling

Brug inkluderende dataindsamlingsteknikker for at sikre, at alle dele af befolkningen er tilstrækkeligt repræsenteret i datasættet. Det kan indebære stratificeret sampling eller oversampling af minoritetsgrupper for at afhjælpe ubalancer.

Rensning af data

Rens og forarbejd dataene omhyggeligt for at fjerne eventuelle skævheder eller uoverensstemmelser. Dette omfatter identifikation og korrektion af fejl, håndtering af manglende værdier og standardisering af dataformater.

Registrering af bias

Implementer teknikker til at opdage og afbøde bias i data. Det kan indebære, at man gennemfører bias-audits eller bruger automatiserede værktøjer til at identificere forskelle på tværs af forskellige demografiske grupper.

Retfærdig funktionsudvikling

Overvej konsekvenserne af valg af funktioner og teknik for retfærdighed. Undgå at bruge funktioner, der kan fastholde stereotyper eller diskrimination, og stræb efter funktioner, der er relevante og ikke-diskriminerende.

Gennemsigtighed og dokumentation

Dokumenter dataindsamling og forbehandlingsprocedurer for at skabe gennemsigtighed og ansvarlighed. Dette omfatter dokumentation af alle beslutninger, der er truffet under datarensning *eller feature engineering*, som kan påvirke retfærdigheden.

Regelmæssig overvågning

Klik på kort for at vende det

sikre retfærdighed over tid.

Retfærdighed i dataindsamling og forbehandling

Forskellige datakilder

Indsaml data fra forskellige kilder for at sikre en omfattende repræsentation af befolkningen. Det er med til at forhindre underrepræsentation eller skævvridning af specifikke demografiske grupper.

Inkluderende dataindsamling

Brug inkluderende dataindsamlingsteknikker for at sikre, at alle dele af befolkningen er tilstrækkeligt repræsenteret i datasættet. Det kan indebære stratificeret sampling eller oversampling af minoritetsgrupper for at afhjælpe ubalancer.

Rensning af data

Rens og forarbejd dataene omhyggeligt for at fjerne eventuelle skævheder eller uoverensstemmelser. Dette omfatter identifikation og korrektion af fejl, håndtering af manglende værdier og standardisering af dataformater.

Registrering af bias

Implementer teknikker til at opdage og afbøde bias i data. Det kan indebære, at man gennemfører bias-audits eller bruger automatiserede værktøjer til at identificere forskelle på tværs af forskellige demografiske grupper.

Retfærdig funktionsudvikling

Overvej konsekvenserne af valg af funktioner og teknik for retfærdighed. Undgå at bruge funktioner, der kan fastholde stereotyper eller diskrimination, og stræb efter funktioner, der er relevante og ikke-diskriminerende.

Gennemsigtighed og dokumentation

Dokumenter dataindsamling og forbehandlingsprocedurer for at skabe gennemsigtighed og ansvarlighed. Dette omfatter dokumentation af alle beslutninger, der er truffet under datarensning eller feature engineering, som kan påvirke retfærdigheden.

Regelmæssig overvågning

Overvåg løbende dataindsamlingen og forbehandlingen for at identificere og håndtere eventuelle nye skævheder eller forskelle. Opdater regelmæssigt datasættet og forbehandlingsteknikkerne for at sikre retfærdighed over tid.

FREMTIDIGE TENDENSER OG NYE UDFORDRINGER

BILLEDKILDE | Genereret af DALL-E

Fremtidige tendenser og nye udfordringer

Forklarlighed og forståelighed

Klik på kort for at vende det

Adversarial-angreb og forsvar

Klik på kort for at vende det

Retfærdighed på tværs af flere dimensioner

Klik på kort for at vende det

Algoritmisk afhjælpning af bias

Klik på kort for at vende det

Etiske overvejelser og styring

Klik på kort for at vende det

Menneskecentreret design

Klik på kort for at vende det

Retfærdighed i nye teknologier

Klik på kort for at vende det

Fremtidige tendenser og nye udfordringer

Forklarlighed og forståelighed

Efterhånden som AI-systemerne bliver mere komplekse, er der en stigende efterspørgsel efter gennemsigtighed og fortolkningsmuligheder. Fremtidige tendenser kan fokusere på at udvikle teknikker og standarder til at forklare AI-beslutninger og gøre dem mere forståelige for interessenter.

Adversarial-angreb og forsvar

Klik på kort for at vende det

Retfærdighed på tværs af flere dimensioner

Klik på kort for at vende det

Algoritmisk afhjælpning af bias

Klik på kort for at vende det

Etiske overvejelser og styring

Klik på kort for at vende det

Menneskecentreret design

Klik på kort for at vende det

Retfærdighed i nye teknologier

Klik på kort for at vende det

Fremtidige tendenser og nye udfordringer

Forklarlighed og forståelighed

Adversarial-angreb og forsvar

Adversarial-angreb, hvor ondsindede aktører manipulerer AI-systemer ved subtilt at ændre inputdata, udgør en betydelig trussel mod retfærdighed og robusthed. Fremtidig forskning kan udforske robuste forsvar og modforanstaltninger mod kontradiktoriske angreb for at forbedre AI-systemers pålidelighed og retfærdighed.

Retfærdighed på tværs af flere dimensioner

Klik på kort for at vende det

Algoritmisk afhjælpning af bias

Klik på kort for at vende det

Etiske overvejelser og styring

Klik på kort for at vende det

Menneskecentreret design

Klik på kort for at vende det

Retfærdighed i nye teknologier

Klik på kort for at vende det

Fremtidige tendenser og nye udfordringer

Forklarlighed og forståelighed

Adversarial-angreb og forsvar

Retfærdighed på tværs af flere dimensioner

Retfærdighed i AI er mere end demografiske attributter og kan omfatte andre dimensioner såsom socioøkonomisk status, handicap eller sprogfærdigheder. Fremtidige tendenser kan indebære udvikling af flerdimensionelle fairnessmålinger og -algoritmer til at håndtere intersektionelle bias.

Algoritmisk afhjælpning af bias

Klik på kort for at vende det

Etiske overvejelser og styring

Klik på kort for at vende det

Menneskecentreret design

Klik på kort for at vende det

Retfærdighed i nye teknologier

Klik på kort for at vende det

Fremtidige tendenser og nye udfordringer

Forklarlighed og forståelighed

Adversarial-angreb og forsvar

Retfærdighed på tværs af flere dimensioner

Algoritmisk afhjælpning af bias

Fremskridt inden for algoritmiske teknikker til reduktion af bias vil fortsat være et fokuspunkt for forskning, der fokuserer på at udvikle mere retfærdige algoritmer og modeller, der reducerer bias på tværs af forskellige beslutningssammenhænge.

Etiske overvejelser og styring

Klik på kort for at vende det

Menneskecentreret design

Klik på kort for at vende det

Retfærdighed i nye teknologier

Klik på kort for at vende det

Fremtidige tendenser og nye udfordringer

Forklarlighed og forståelighed

Adversarial-angreb og forsvar

Retfærdighed på tværs af flere dimensioner

Algoritmisk afhjælpning af bias

Etiske overvejelser og styring

De etiske konsekvenser af AI-retfærdighed og bias vil fortsat være et centralt område med stigende vægt på etisk AI-udviklingspraksis og styringsrammer. Fremtidige tendenser kan indebære etablering af lovgivningsmæssige retningslinjer og standarder for at sikre retfærdighed og ansvarlighed i AI-systemer.

Menneskecentreret design

Klik på kort for at vende det

Retfærdighed i nye teknologier

Klik på kort for at vende det

Fremtidige tendenser og nye udfordringer

Forklarlighed og forståelighed

Adversarial-angreb og forsvar

Retfærdighed på tværs af flere dimensioner

Algoritmisk afhjælpning af bias

Etiske overvejelser og styring

Menneskecentreret design

Fremtidige AI-systemer kan prioritere menneskecentrerede designprincipper og inddrage brugernes feedback og præferencer for at øge retfærdigheden og anvendeligheden. Det kan indebære co-design af AI-systemer med forskellige interessenter for at sikre inklusion og retfærdighed.

Retfærdighed i nye teknologier

Klik på kort for at vende det

Fremtidige tendenser og nye udfordringer

Forklarlighed og forståelighed

Adversarial-angreb og forsvar

Retfærdighed på tværs af flere dimensioner

Algoritmisk afhjælpning af bias

Etiske overvejelser og styring

Menneskecentreret design

Retfærdighed i nye teknologier

I takt med at AI fortsætter i nye teknologier som f.eks. autonome køretøjer, AI i sundhedssektoren og ansigtsgenkendelsessystemer, bliver det stadig vigtigere at tage fat på retfærdighed og bias. Fremtidige tendenser kan indebære tilpasning af retfærdighedsprincipper til nye anvendelser og teknologiske fremskridt.

KONKLUSION

BILLEDKILDE | Genereret af DALL-E

KONKLUSION

1. **Forståelse af kompleksitet**: AI-retfærdighed og bias er mangefacetterede udfordringer, der kræver omfattende strategier og løbende opmærksomhed.
2. **Strategier for handling**: Fra mangfoldig repræsentation i dataindsamling til algoritmisk afhjælpning af bias er der en række strategier til rådighed for at håndtere og afhjælpe bias i AI-systemer.
3. **Forudse nye udfordringer**: Efterhånden som AI fortsætter med at udvikle sig, vil det være afgørende at være på forkant med nye tendenser og problemer som f.eks. fjendtlige angreb, flerdimensionel retfærdighed og etiske overvejelser.
4. **Mod en mere retfærdig fremtid**: Ved at prioritere menneskecentreret design, gennemsigtighed og ansvarlighed kan vi arbejde hen imod at bygge AI-systemer, der fremmer retfærdighed, inklusivitet og tillid.

TAK

Medfinansieret af Den Europæiske Union

Finansieret af Den Europæiske Union. Synspunkter og holdninger, der kommer til udtryk, er udelukkende forfatterens/forfatternes og er ikke nødvendigvis udtryk for Den Europæiske Unions eller Det Europæiske Forvaltningsorgan for Uddannelse og Kulturs (EACEA) officielle holdning. Hverken den Europæiske Union eller EACEA kan holdes ansvarlig herfor.

2022-1-ES01-KA220-HED-000085257