



# Etisk AI-mikrocertifikat

HÆFTE

CU1 | Hvad er algoritmisk bias?

# Hvordan bruger man denne flipbook?

Dette dokument er interaktivt. I hele dokumentet finder du links til yderligere information.



Knap, der fører dig til begyndelsen af dokumentet. Dette ikon vises i øverste højre hjørne af siderne.



Når du ser denne pil, betyder det, at du har en **interaktiv farvetekst** at klikke på, som er forbundet med et eksternt link.

**ANSVARSFRAKRIVELSE:** Bemærk, at vi ikke kan garantere den fortsatte tilgængelighed af eksternt indhold, f.eks. videoer, da de kan ændres eller fjernes af deres forfattere eller værtsplatforme.

# Indeks

Klik på menuen

**01. Introduktion**

**02. Kursets indhold og forventede resultater**

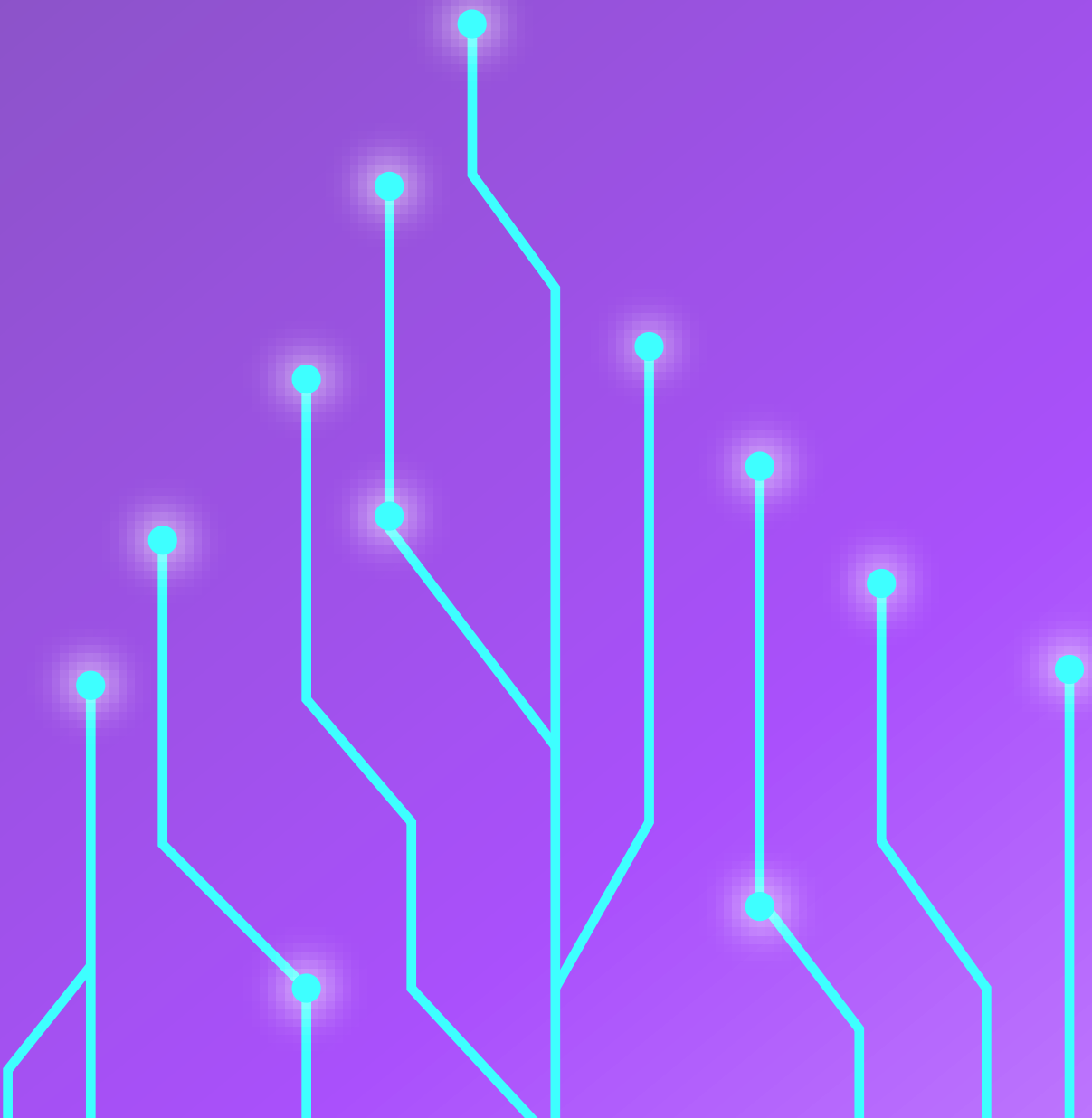
**03. Hvad er algoritmisk bias?**

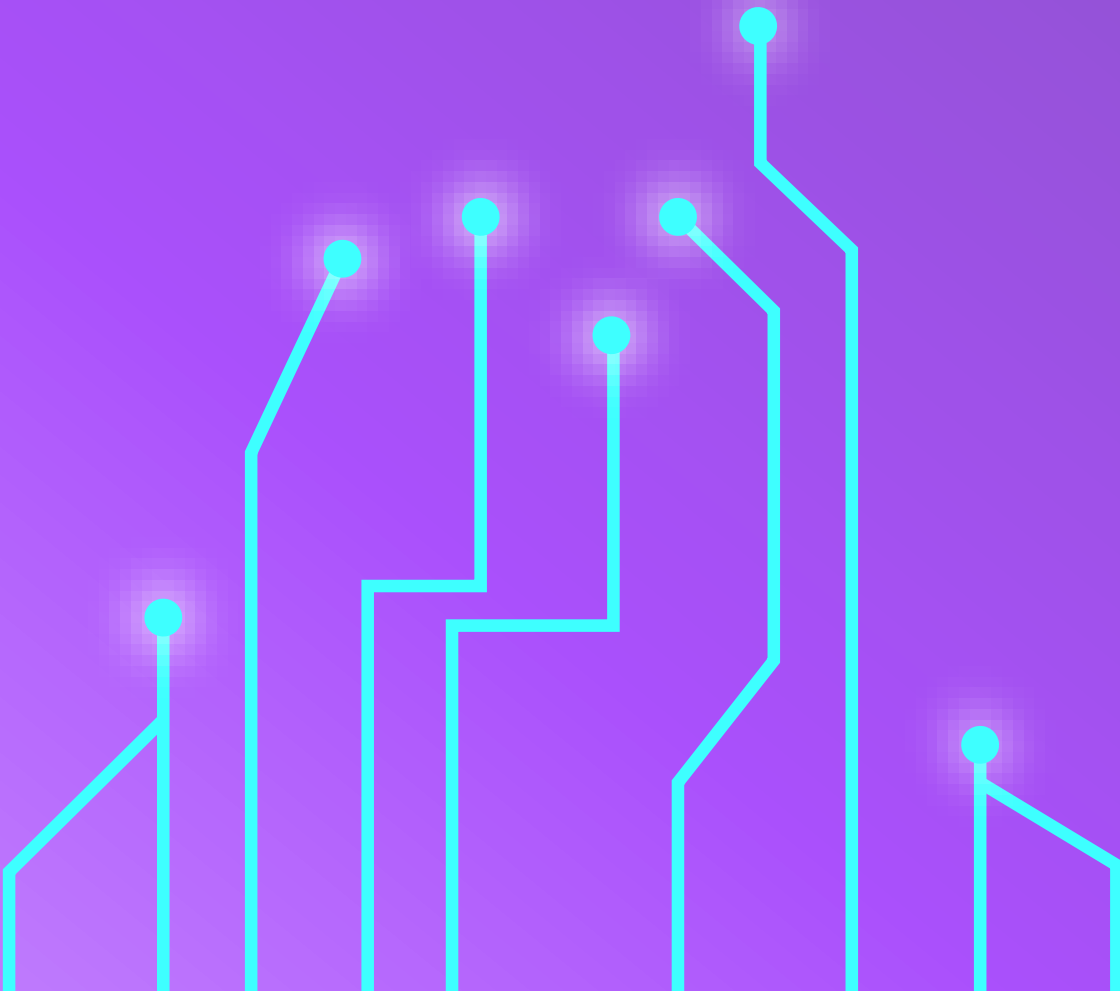
**04. Definition af algoritmisk bias**

**05. Forståelse af bias i AI-systemer**

# 01. Introduktion

CU1 | Hvad er algoritmisk bias?





## 01. Introduktion

I det hurtigt udviklende landskab af kunstig intelligens (AI) er det altafgørende at sikre etisk ansvarlig udvikling og brug af AI-teknologier. Dette hæfte fungerer som en omfattende guide til Ethical AI microcredential med fokus på seks kompetenceenheder, der er designet til at udstyre dig med den viden og de færdigheder, der er nødvendige for at navigere i de etiske kompleksiteter ved AI.

Når du går i gang med denne rejse, vil du udforske seks forskellige kompetenceenheder, der hver især behandler vigtige aspekter af etisk AI-udvikling og -anvendelse. Fra forståelse af algoritmisk bias til fremme af gennemsigtighed og opretholdelse af menneskerettigheder, er disse kompetenceenheder designet til at give dig de nødvendige værktøjer til at navigere i de etiske udfordringer, der er forbundet med AI-teknologier.

I løbet af dette hæfte vil du dykke ned i følgende kompetenceenheder (CU fra nu af):

- CU1 - Hvad er algoritmisk bias?
- CU2 - Non-maleficence
- CU3 - Ansvarlighed
- CU4 - Gennemsigtighed
- CU5 - Menneskerettigheder og retfærdighed
- CU6 - AI-etik, en praktisk tilgang



Hver enhed giver dig en dybere forståelse af vigtige etiske principper og praksisser inden for AI, sammen med praktisk indsigt og eksempler fra den virkelige verden, der styrker din læring.

Uanset om du er en elev, studerende, en professionel eller en AI-entusiast, tilbyder dette hæfte en værdifuld ressource til at udvide din viden og ekspertise inden for etisk AI. Vi inviterer dig med på rejsen, hvor vi udforsker de etiske dimensioner af kunstig intelligens og arbejder på at skabe en mere ansvarlig og retfærdig fremtid.

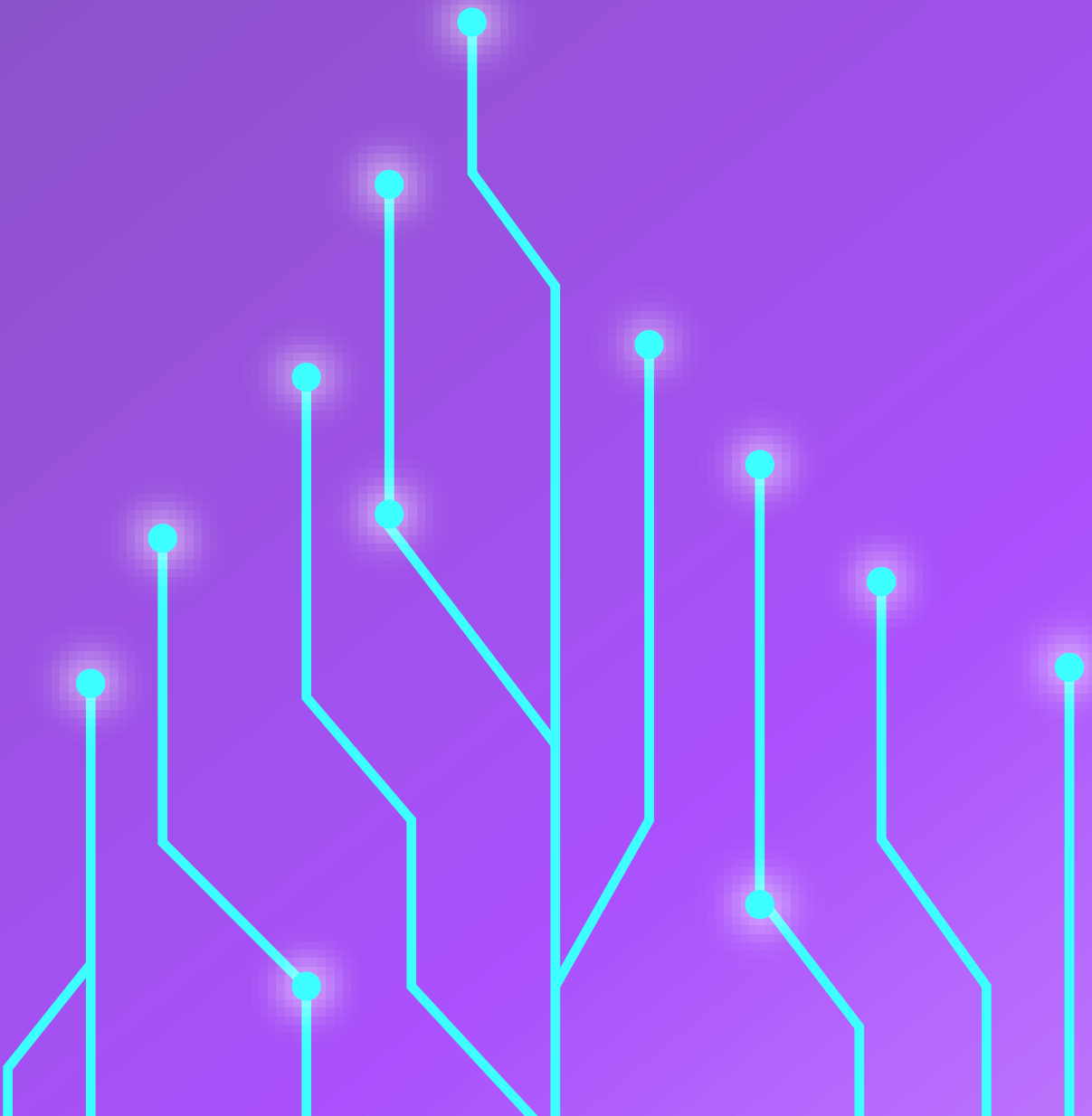
Tak, fordi du har valgt dette hæfte som din guide til etisk AI-udvikling og -praksis.

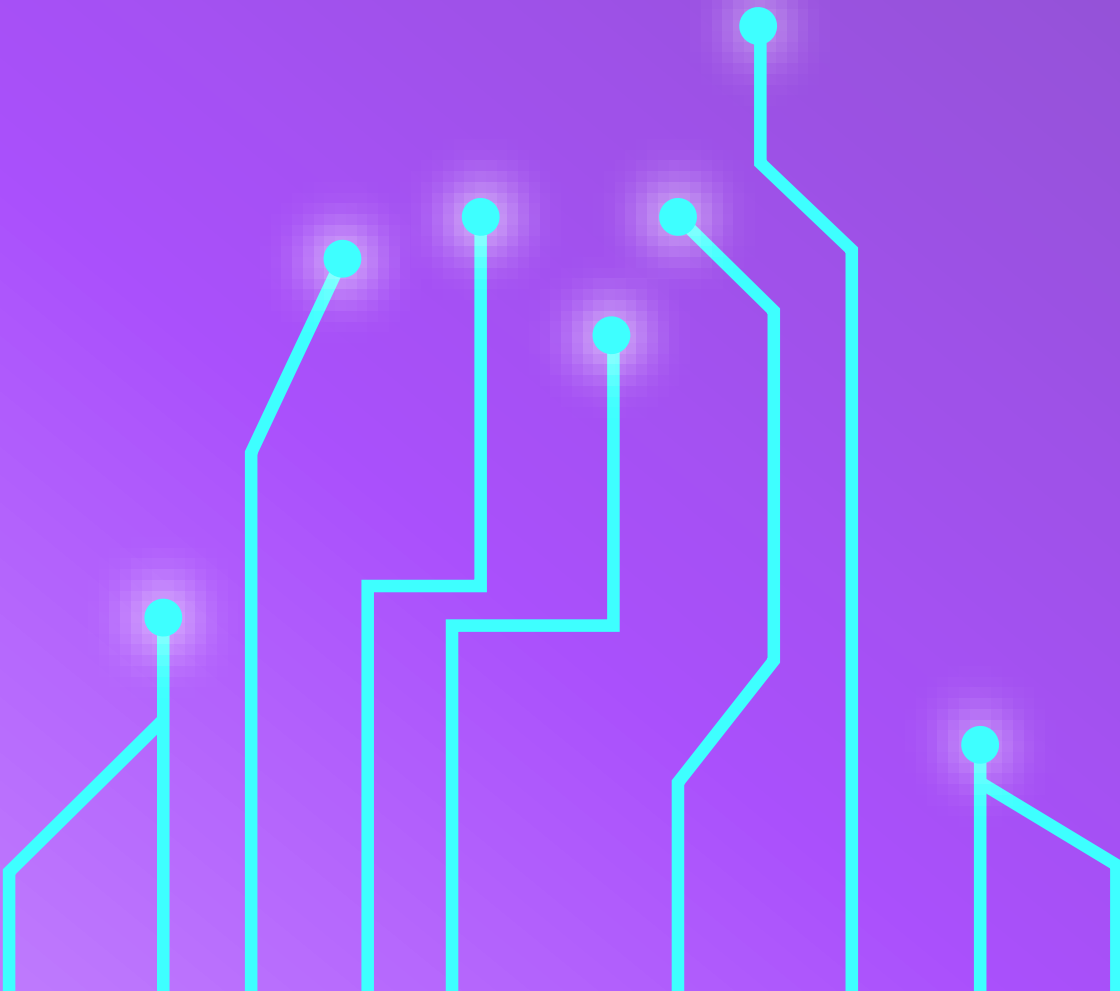
Lad os begynde denne transformative rejse sammen!

CHARLIE-projektets team

# 02. Kursets indhold og forventede resultater

CU1 | Hvad er algoritmisk bias?





## 02. Kursets indhold og forventede resultater

"Ethical AI Microcredential" på EQF4-niveau er designet til at opnå følgende resultater:

1. Etablere en grundlæggende forståelse af algoritmisk bias og udforske dens oprindelse og konsekvenser for individer og samfund.
  - Dykke ned i definitionen, kilderne og manifestationerne af algoritmisk bias.
  - Analysere de samfundsmæssige og individuelle konsekvenser af forudindtagede algoritmer.
2. Dyrke bevidstheden om og anvendelsen af det etiske princip om ikke-skadevirkning i AI-udvikling.
  - Vurdere de risici og skader, der er forbundet med forudindtagede algoritmer.
  - Udvikle strategier til at afbøde skader og fremme etisk AI-udvikling.
3. Værdsætte betydningen af ansvarlighed i AI-systemer og undersøge relevante juridiske og etiske rammer.
  - Undersøge forskellige interessenters roller i forbindelse med AI-ansvarlighed.
  - Lære om de bedste praksisser for at fremme ansvarlighed i AI-udvikling.



4. Få indsigt i begrebet **gennemsigtighed i AI-systemer** og dets centrale rolle i algoritmisk beslutningstagning.
  - Udforske metoder og værktøjer til at øge gennemsigtigheden i AI.
  - Forstå udfordringerne og begrænsningerne ved at gøre komplekse algoritmer mere forståelige.
5. Udforsk **krydsfeltet mellem AI, menneskerettigheder og retfærdighed** og deres konsekvenser for etisk AI-udvikling.
  - Vurdere indvirkningen af forudindtagede algoritmer på menneskerettighederne, herunder ikke-diskrimination, privatliv og ytringsfrihed.
  - Udvikle strategier for at sikre retfærdighed i udviklingen og implementeringen af AI.
6. Anvend etiske principper i udvikling og implementering af AI gennem **praktiske tilgange og scenarier fra den virkelige verden**.
  - Undersøge forskellige etiske rammer og retningslinjer og deres anvendelse på AI-systemer.
  - Forstå vigtigheden af inddragelse af interessenter, tværfagligt samarbejde og etiske AI-udviklingsprocesser.



Når dette kursus afsluttes, vil deltagerne have en holistisk forståelse af algoritmiske bias, deres sektorspecifikke konsekvenser og værktøjer og strategier til at håndtere disse. Denne viden ruster fagfolk og akademikere/studerende inden for algoritmedrevne områder til at bidrage til lige og retfærdige resultater i en datadrevet verden.

Microcredential-kurset er struktureret omkring 6 kompetenceenheder (CU'er), der hver især er designet til at udstyre deltagerne med den viden og de færdigheder, der er nødvendige for at navigere i udfordringerne og mulighederne inden for algoritmisk bias.

**CU1 - Hvad er algoritmisk bias?** I denne enhed vil eleverne udforske begrebet algoritmisk bias og dets forskellige manifestationer. Den dækker dets definition, årsager og samfundsmæssige konsekvenser. Eleverne vil analysere oprindelse af bias, kilder inden for algoritmer og potentielle påvirkninger af individer og samfund.

**CU2 - Non-maleficence:** Denne enhed dykker ned i det etiske princip om non-maleficence, der prioriterer at undgå skade i udvikling og anvendelse af AI. Deltagerne vil udforske de iboende risici og skader, der er forbundet med forudindtagede algoritmer, samtidig med at de afdækker strategier til at afbøde disse risici og fremme etisk AI-praksis.



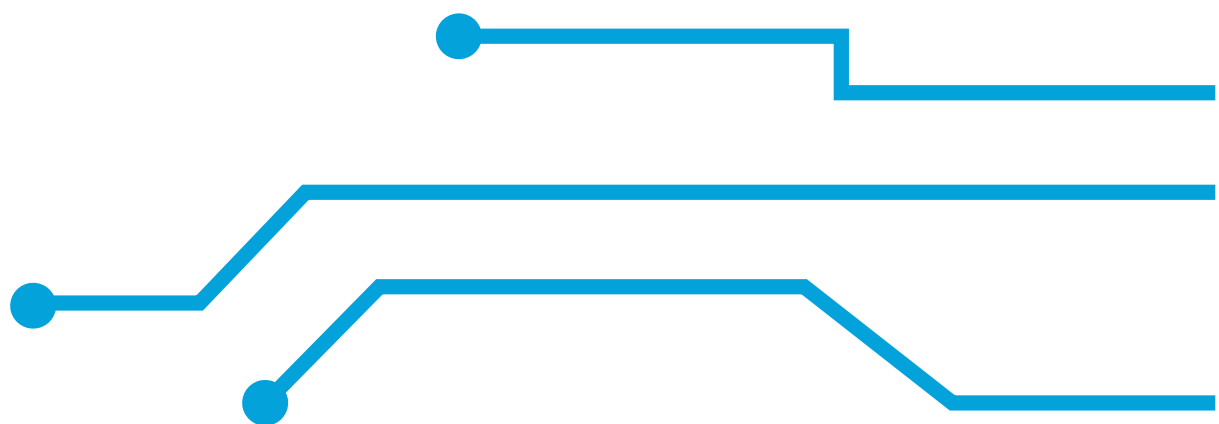
**CU3 - Ansvarlighed:** I denne enhed dykker eleverne ned i det kritiske område af ansvarlighed inden for udvikling og brug af AI. Deltagerne uddyber nødvendigheden af at udvikle klare ansvarslinjer og udforske de juridiske og etiske rammer for ansvarlighed. Derudover undersøger læseplanen de forskellige interessenters roller og dykker ned i den bedste fremgangsmåde, der sikrer ansvarlighed i bestræbelserne i forbindelse med udvikling af AI.

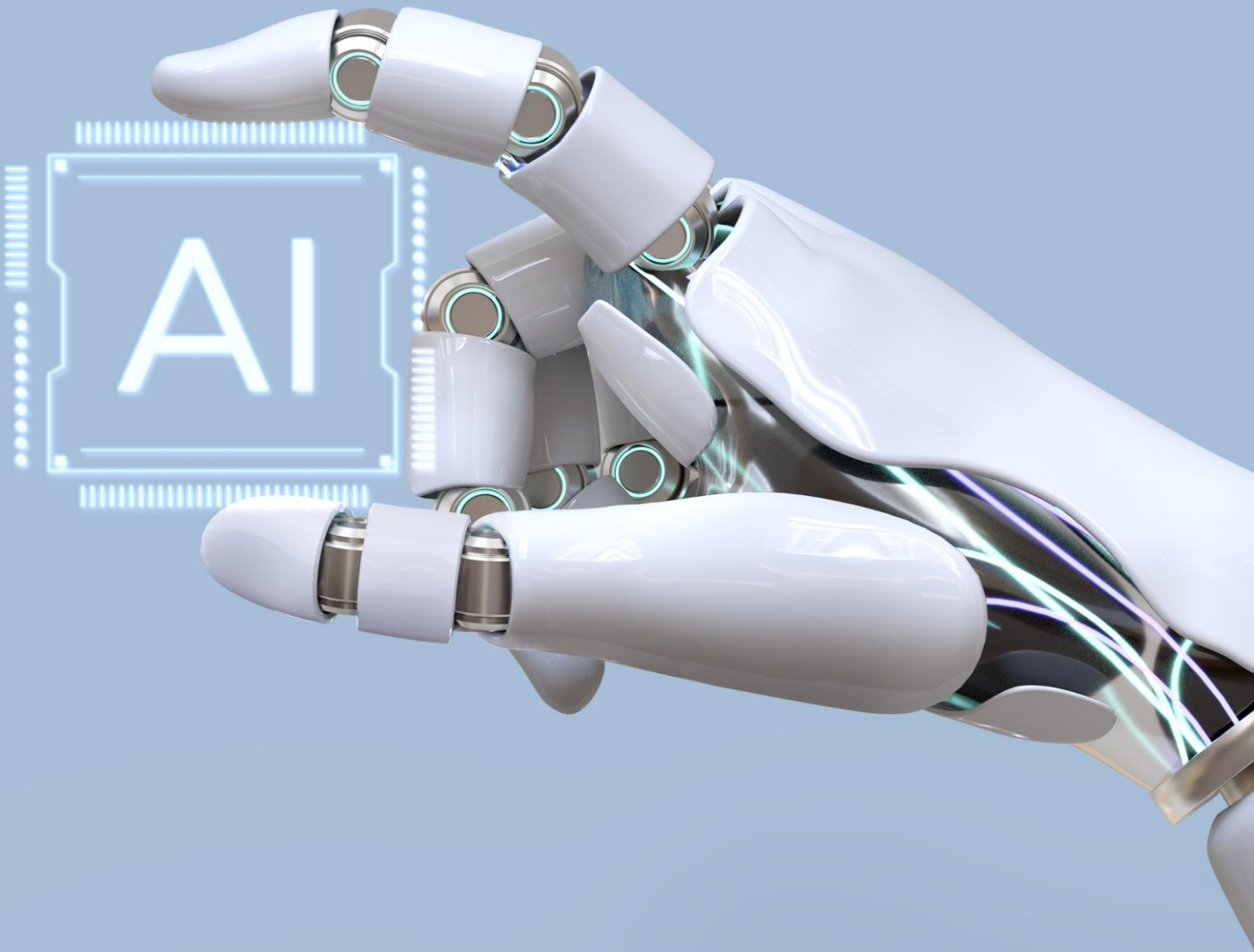
**CU4 - Gennemsigtighed:** Denne enhed belyser betydningen af gennemsigtighed i AI-systemer og understreger værdierne af åbenhed, kommunikation og transparens i algoritmisk beslutningstagning. Deltagerne vil beskæftige sig med teknikker og ressourcer, der har til formål at øge gennemsigtigheden i AI, samtidig med at de kæmper med de iboende udfordringer og begrænsninger, der er forbundet med at gøre udviklede algoritmer forståelige.

**CU5 - Menneskerettigheder og retfærdighed:** I enheden Menneskerettigheder og retfærdighed vil eleverne udforske krydsfeltet mellem AI, menneskerettigheder og retfærdighed. De vil undersøge, hvordan forudindtagede algoritmer kan påvirke menneskerettighederne, herunder retten til ikke-diskrimination, privatliv og ytringsfrihed. Eleverne vil også lære om strategier til at sikre retfærdighed og lighed i udviklingen og implementeringen af AI.

**CU6 - AI-etik, en praktisk tilgang:** Denne enhed lægger vægt på den pragmatiske anvendelse af etiske principper i hele udviklingen og implementeringen af AI. Deltagerne dykker ned i forskellige etiske rammer og retningslinjer og får indsigt i deres anvendelse i den virkelige verden med udgangspunkt i AI-scenarier. Derudover understreger enheden betydningen af interessenternes engagement, tværfagligt samarbejde og integration af etiske AI-udviklingsprocesser for at fremme ansvarlig AI-innovation.

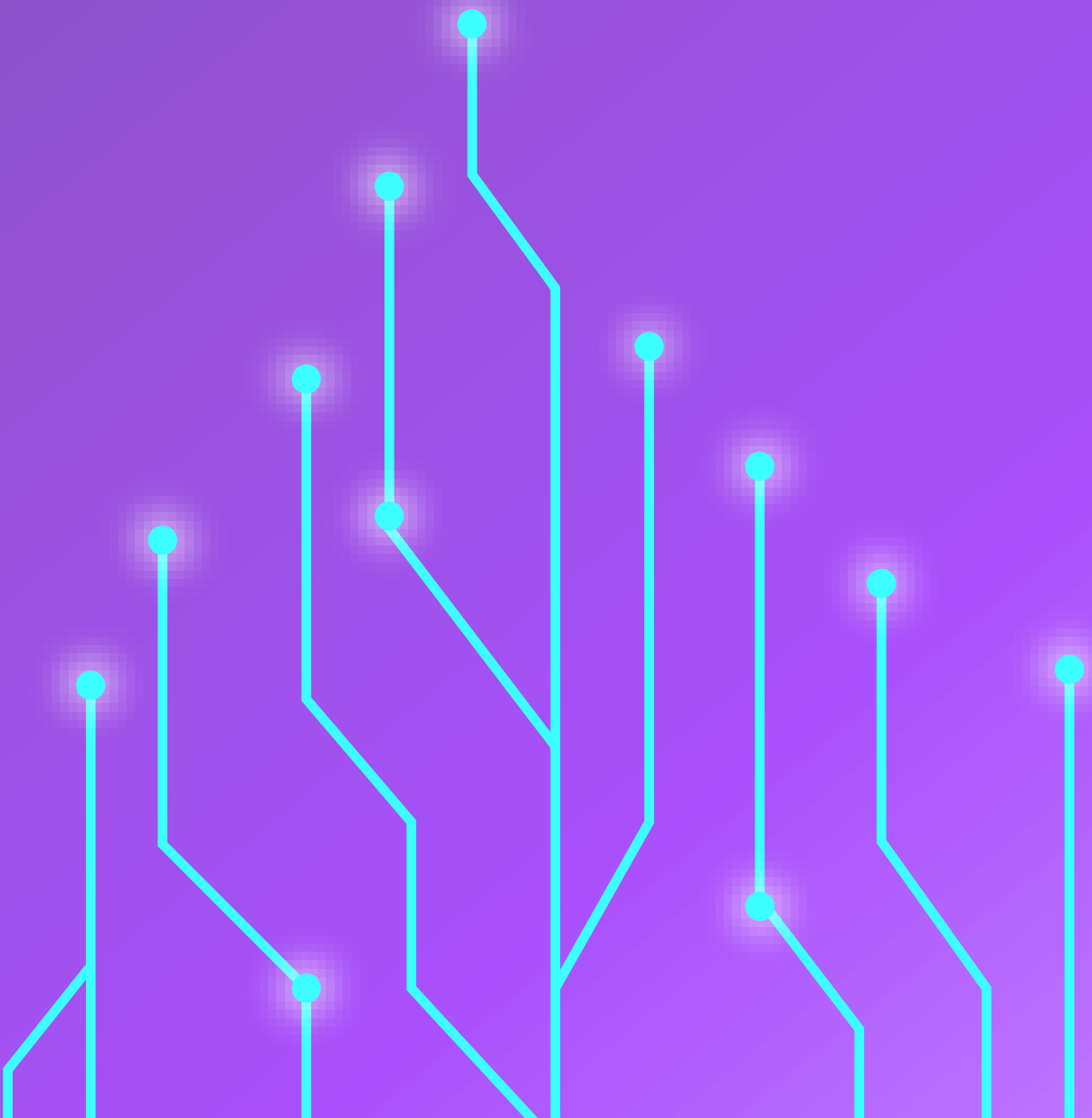
Det efterfølgende afsnit uddyber indholdet af hver kompetenceenhed.

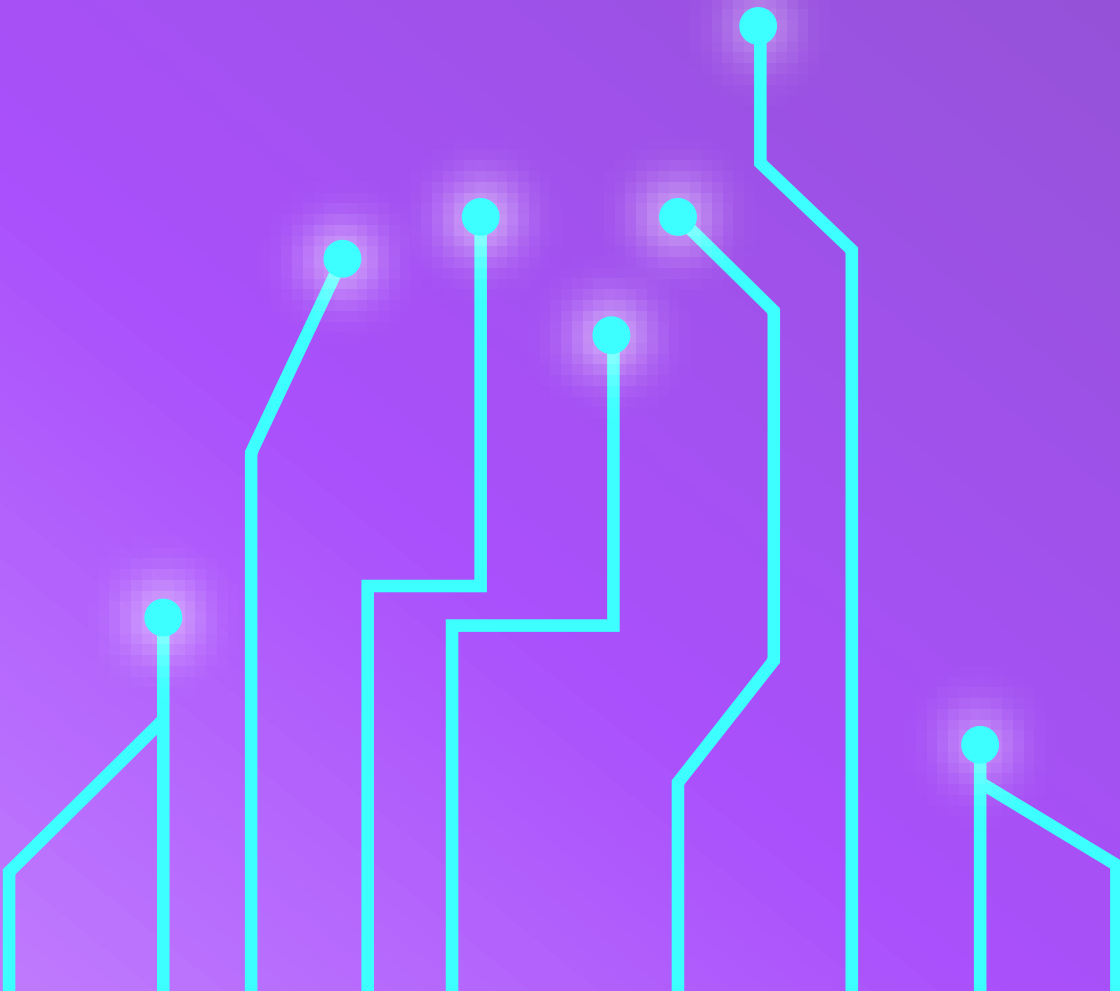




# 03. Hvad er algoritmisk bias?

CU1 | Hvad er algoritmisk bias?





### 03. Hvad er algoritmisk bias?

Algoritmer bruges til at træffe vigtige beslutninger. Men de kan nogle gange være forudindtagede og uretfærdige over for visse grupper af mennesker. Dette er kendt som algoritmisk bias.

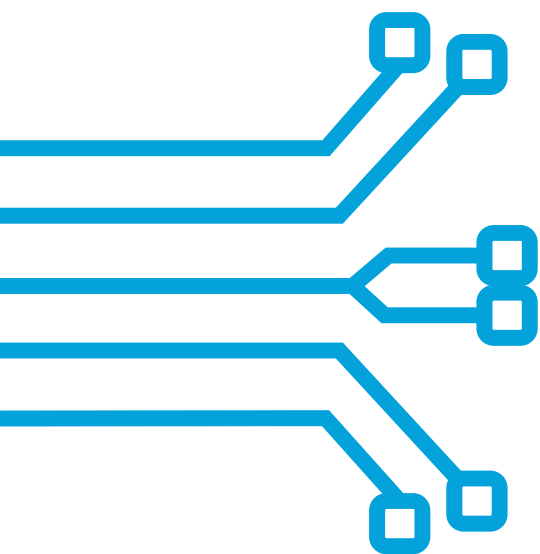
I denne kompetenceenhed vil eleverne lære om algoritmisk bias, dens forskellige former, og hvordan man identificerer den. De vil også udforske årsagerne til bias i algoritmer, herunder indvirkningen af menneskelig bias på beslutningstagning. Derudover vil eleverne undersøge de potentielle konsekvenser af forudindtagede algoritmer for enkeltpersoner og samfundet, hvilket kan føre til diskrimination og uretfærdig behandling. Ved afslutningen af dette forløb vil de studerende have en bedre forståelse af algoritmisk bias, og hvordan de kan håndtere det i deres fremtidige arbejde.

Vidensresultaterne for denne enhed omfatter:

- **Definition af algoritmisk bias:** Eleverne vil lære om algoritmisk bias og dens årsager, herunder partisk dataindsamling, skæve træningsdata og menneskelig beslutningstagning. Denne viden vil hjælpe dem med at forstå, hvordan bias kan påvirke AI-applikationer, som f.eks. ansigtsgenkendelsessystemer, der fejlidentificerer visse grupper af mennesker.

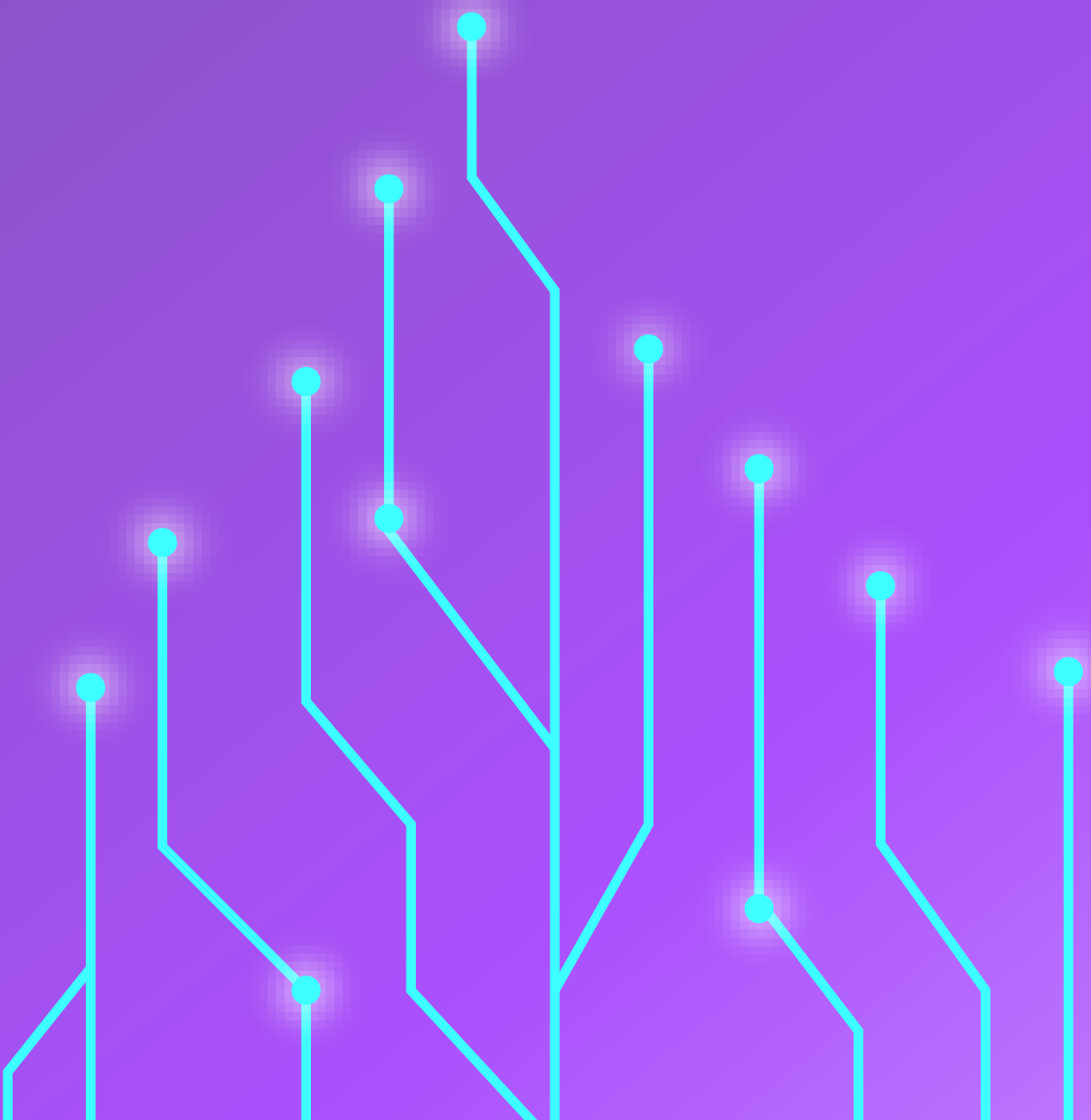


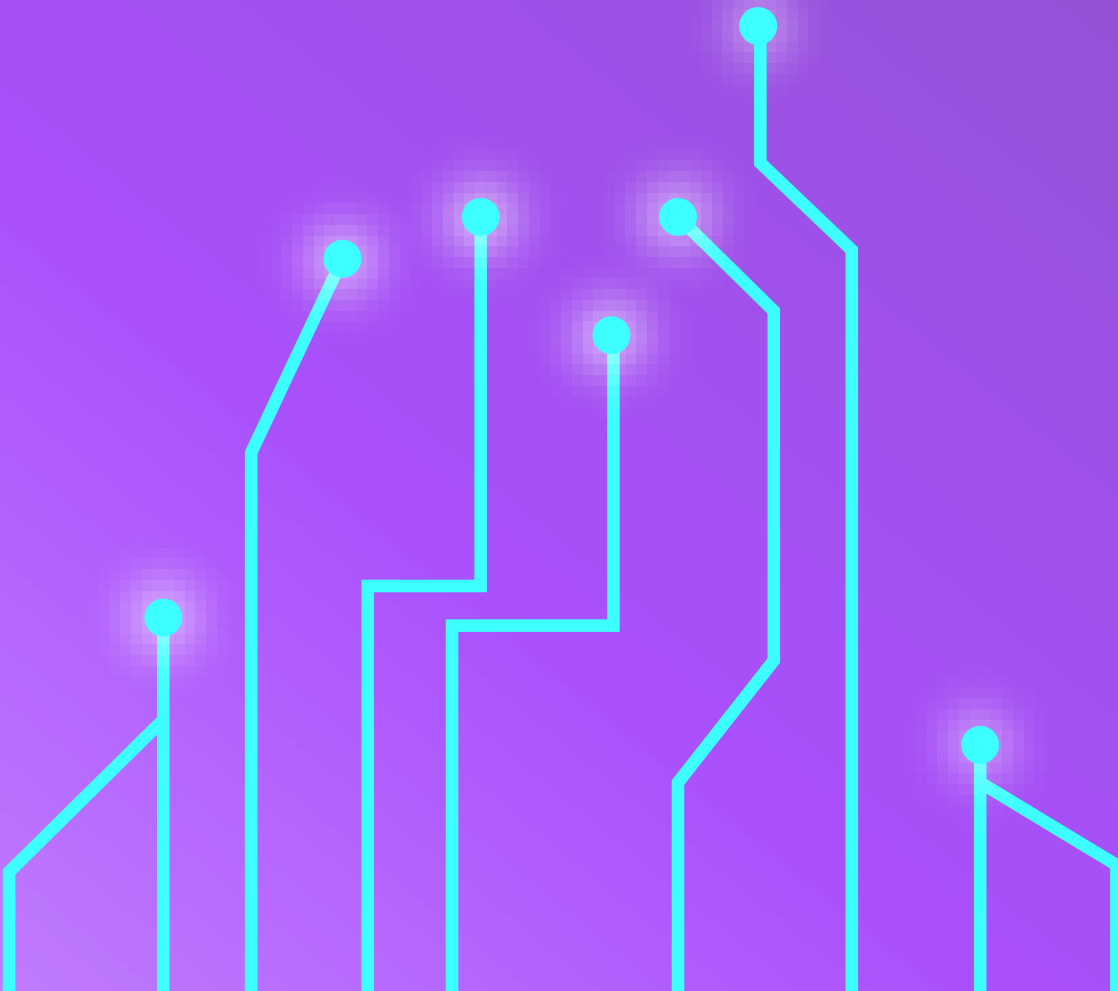
- **Identificering af typer af algoritmiske bias:** Eleverne vil lære om algoritmiske bias, herunder datadrevne, modeldrevne og menneskedrevne bias. De vil forstå, hvordan disse bias kan forårsage uretfærdighed i AI-systemer. For eksempel kan datadrevet bias skyldes ikke-repræsentative træningsdata, hvilket fører til forudindtagede forudsigelser inden for områder som kreditscoring eller screening af jobansøgere.
- **Den virkelige verdens konsekvenser af algoritmisk bias:** På dette kursus lærer eleverne om konsekvenserne af algoritmisk bias i forskellige sektorer som f.eks. sundhed, finans og strafferet. De vil forstå behovet for at minimere algoritmisk bias i AI-systemer for at fremme fairness og retfærdighed. Eksempler på forudindtagede AI-systemer, der fører til negative resultater inden for sundhedspleje og strafferet vil blive diskuteret.



# 04. Definition af algoritmisk bias

CU1 | Hvad er algoritmisk bias?





## 04. Definition af algoritmisk bias

Algoritmisk bias er et kritisk aspekt af AI, som har fået opmærksomhed i de senere år. Det er vigtigt at forstå, at det er for alle, der er involveret i udvikling, implementering eller regulering af AI. Lad os definere, hvad algoritmisk bias er, og hvorfor det er afgørende at undersøge.

### > Hvad er algoritmisk bias?

Algoritmisk bias refererer til systematiske fejl eller uretfærdighed i resultaterne af AI-systemer på grund af forskellige faktorer som forvrængede data, fejlbehæftede algoritmer eller menneskelig beslutningstagning. Disse skævheder kan føre til diskriminerende eller uretfærdig behandling af enkeltpersoner eller grupper, fastholde eksisterende sociale uligheder og forstærke stereotyper.

### Hvorfor studere algoritmisk bias?

For at kunne udforske de forskellige former, årsager og konsekvenser af algoritmisk bias, er det vigtigt først at forstå dets definition og betydning. Med denne viden kan vi udstyre os selv med værktøjerne til at identificere, afbøde og forhindre algoritmisk bias i AI-systemer.

- 1. Etiske konsekvenser:** Algoritmisk bias kan resultere i uretfærdig behandling af personer baseret på race, køn, alder eller andre karakteristika, hvilket er i strid med principperne om fairness og retfærdighed.



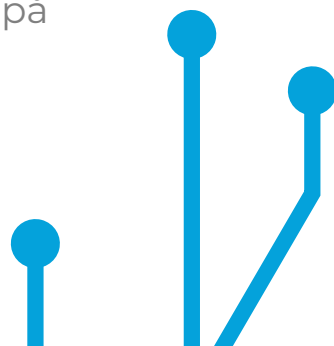
2. **Sociale konsekvenser:** Biased AI-systemer kan forværre samfundsmæssige uligheder og diskrimination og påvirke marginaliserede samfunds adgang til muligheder, ressourcer og tjenester.
3. **Juridiske og lovgivningsmæssige bekymringer:** Efterhånden som AI-teknologier bliver mere udbredte, er der stigende opmærksomhed fra lovgivere og reguleringsorganer på at håndtere algoritmisk bias for at sikre overholdelse af antidiskriminationslove og beskytte enkeltpersoners rettigheder.
4. **Omdømme og tillid:** Organisationer, der anvender forudindtagede AI-systemer, risikerer at skade deres omdømme og miste offentlighedens tillid, hvilket kan have betydelige konsekvenser for deres brandimage og troværdighed på markedet.



## > Faktorer, der bidrager til skævvredne resultater

Flere indbyrdes forbundne faktorer bidrager til fremkomsten af forudindtagede AI-systemer og underminerer deres pålidelighed, retfærdighed og effektivitet. I dette afsnit undersøger vi nogle af de mest almindelige faktorer, der bidrager til forudindtagede resultater i AI-systemer.

- **Forvrængede data:** Biased data, der bruges til at træne AI-systemer, resulterer i algoritmisk bias, som kan føre til diskriminerende resultater. For at afbøde dette skal der foretages omhyggelige overvejelser i forbindelse med dataindsamling og forbehandling, herunder repræsentativ prøveudtagning, biasdetektering og afbødningsalgoritmer og forskelligartet dataforøgelse.
- **Fejlbehæftede algoritmer:** AI-systemer kan have forudindtagede resultater på grund af fejlbehæftede algoritmer, designvalg, modelarkitekturer, optimeringsprocedurer eller inputvariabler. Fairness-bevidst maskinlæring, algoritmisk gennemsigtighed og fortolkningsteknikker kan hjælpe med at afbøde sådanne skævheder.
- **Menneskelige bias:** Bias i AI-systemer kan skyldes ubevidste påvirkninger fra udviklere, dataforskere og beslutningstagere. For at undgå disse bias bør AI-udviklingsteams fokusere på mangfoldighed, etiske retningslinjer og ansvarlighedsmekanismer.





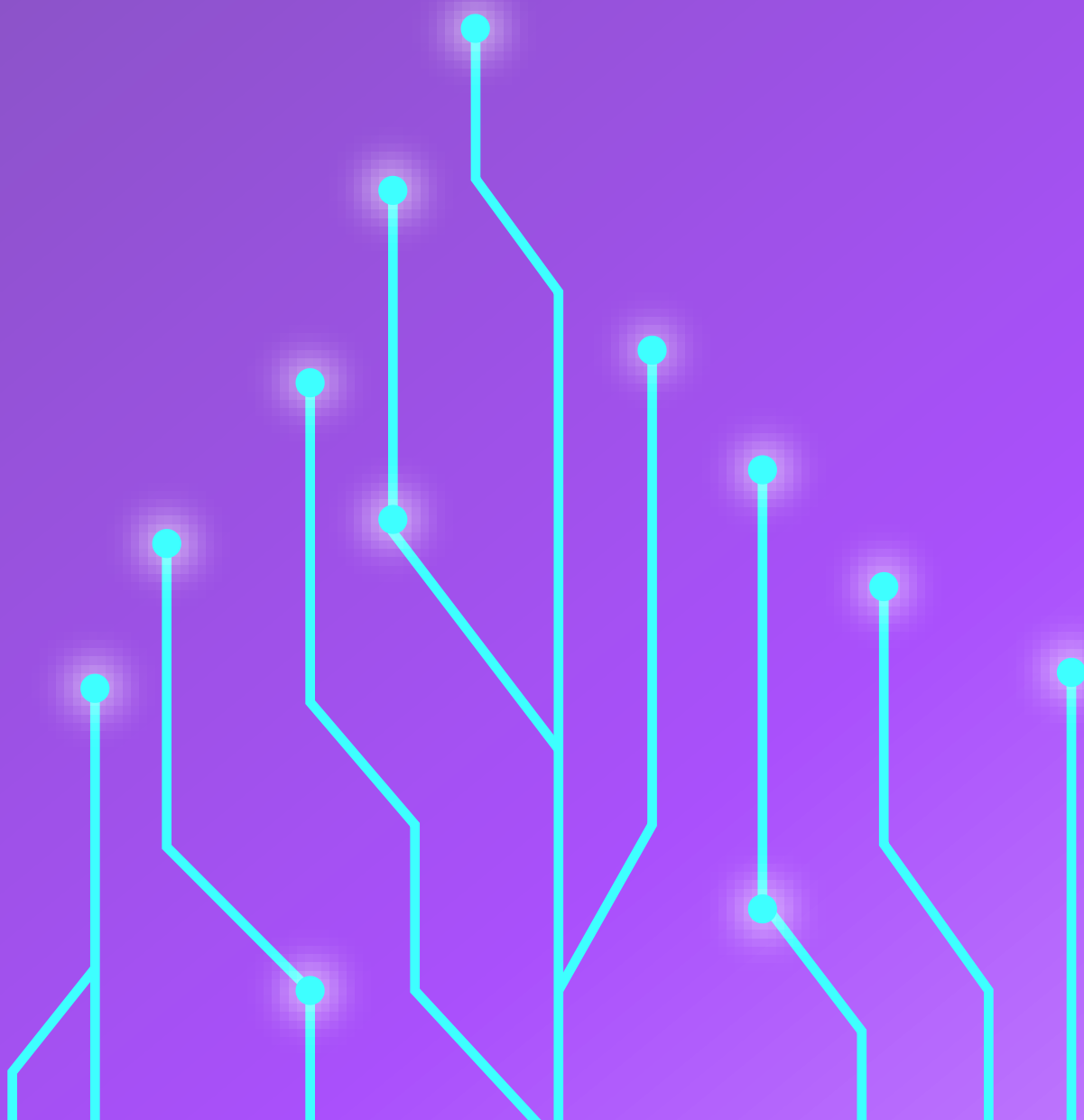
## > **Eksempler på forudindtagede systemer**

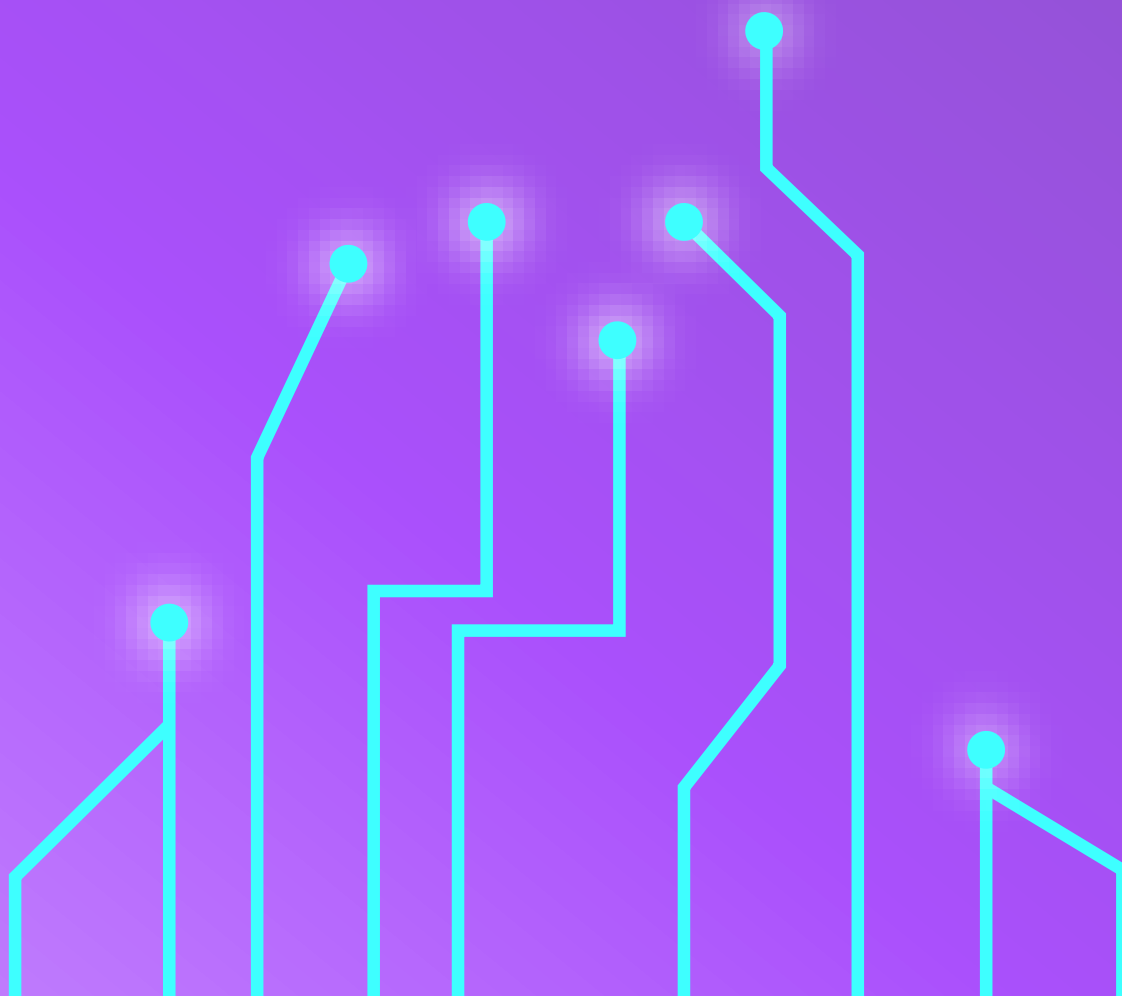
AI-systemer kan være forudindtagede og føre til uretfærdige resultater. Nedenfor er der nogle hurtige eksempler fra den virkelige verden om AI-systemer, der ofte er forudindtagede, og som fremhæver de potentielle konsekvenser af algoritmiske bias. Vi vil udforske dem dybdegående senere i dette kursus.

- **Algoritmer til ansigtsgenkendelse:** Ansigtsgenkendelsesteknologi kan have bias, der fastholder race- eller kønsforskelle, hvilket fører til uretmæssige anholdelser eller overvågning af specifikke grupper. Det er afgørende at tage fat på disse skævheder for at sikre retfærdighed i AI-systemer og genoprette offentlighedens tillid.
- **Forudsigende politialgoritmer:** Predictive policing-algoritmer kan videreføre skævheder i historiske kriminalitetsdata og føre til overpolitisering af visse samfund eller demografiske grupper. Skævvredne algoritmer kan forværre eksisterende forskelle i retshåndhævelsespraksis og give anledning til bekymring om retfærdighed, ansvarlighed og potentiale for diskriminerende resultater i strafferetlige systemer.
- **Automatiserede ansættelsessystemer:** Automatiserede ansættelsessystemer kan fastholde fordomme, føre til diskriminerende praksis og begrænse mangfoldigheden i arbejdsstyrken. Biased algoritmer kan lære mønstre af bias fra historiske data, hvilket resulterer i fortrinsbehandling af visse demografiske grupper. Revision og afhjælpning af bias er afgørende for at sikre fairness, retfærdighed og ansvarlighed i AI-drevne rekrutteringsprocesser.

# 05. Forståelse af bias i AI-systemer

CU1 | Hvad er algoritmisk bias?





## 05. Forståelse af bias i AI-systemer

I dette afsnit vil vi udforske tre typer af bias: **datadrevet**, **modeldrevet** og **menneskedrevet**.

Disse bias kan påvirke nøjagtigheden og troværdigheden af AI-systemer, og at forstå dem er det første skridt i retning af at forebygge dem.

### > **Datadrevet forudindtagethed**

Hvad er datadrevet bias?

Datadrevet bias refererer til bias, der opstår på grund af karakteristika eller fordeling af de træningsdata, der bruges til at udvikle maskinlæringsmodeller.

Skævvredne træningsdata kan afspejle historiske uligheder, samfundsmæssige fordomme eller systemisk diskrimination, hvilket fører til skæve repræsentationer af visse demografiske grupper eller underrepræsentation af andre.

Forståelse af datadrevet bias er afgørende for at erkende, hvordan forudindtaget træningsdata kan fastholde og forværre eksisterende stereotyper, uligheder og diskriminerende praksis i AI-systemer.



## Årsager til datadrevet bias

### 1. Ufuldstændig eller forudindtaget prøveudtagning:

Træningsdatasæt kan mangle mangfoldighed eller ikke repræsentere visse demografiske grupper tilstrækkeligt, hvilket fører til skæve repræsentationer og forudindtagede modelforudsigelser.

### 2. Historiske fordomme:

Træningsdata kan afspejle historiske uligheder eller systemiske skævheder i samfundet, hvilket viderefører diskriminerende resultater i AI-systemer.

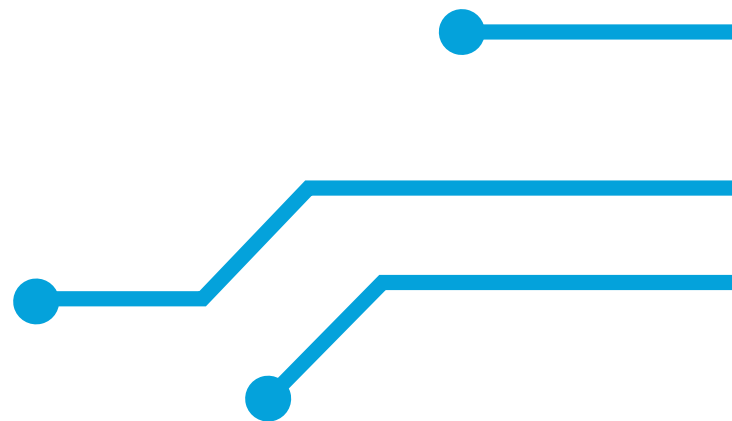
### 3. Mærkningsforstyrrelser:

Skæv eller subjektiv mærkningspraksis kan indføre skævheder i træningsdata, påvirke modelforudsigelser og forstærke eksisterende stereotyper.



## Eksempler på datadrevet bias

- 1. Biased ansigtsgenkendelse:** Ansigtsgenkendelsesalgoritmer, der er trænet på ukorrekte datasæt, kan udvise racemæssige eller køns-mæssige bias, hvilket fører til fejlidentifikation og diskrimination af visse demografiske grupper.
- 2. Kønsbias i sprogmodeller:** Sprogmodeller, der er trænet på forudindtaget tekst, kan generere kønsstereotypisk eller diskriminerende sprog, der afspejler og viderefører samfundsmæssige fordomme.
- 3. Racemæssig bias i forudsigende politiarbejde:** Predictive policing-algoritmer, der er trænet på forudindtagede kriminalitetsdata, kan være uforholdsmæssigt rettet mod minoritetssamfund og forværre racemæssige forskelle i retshåndhævelsen.





## Indvirkning af datadrevet bias

- 1. Forstærkning af stereotyper:** Skævvredne træningsdata kan forstærke eksisterende stereotyper og fordomme og dermed fastholde diskrimination og ulighed i AI-systemer.
- 2. Forstærkning af ulighed:** Datadrevet bias kan forværre eksisterende uligheder og forskelle, hvilket fører til uretfærdig behandling og diskriminerende resultater for marginaliserede grupper.
- 3. Erosion af tillid:** Fordomsfulde AI-systemer underminerer tilliden til teknologi og forværrer bekymringer om retfærdighed, ansvarlighed og gennemsigtighed.

Datadrevet bias er en betydelig udfordring for fair og retfærdige AI-systemer. Ved at forstå dens årsager og konsekvenser kan interessenter tage proaktive skridt til at afbøde bias i træningsdata og fremme inklusion i AI.



## > Model-drevet bias

Hvad er modeldrevet bias?

Modeldrevet bias refererer til bias, der opstår som følge af design, struktur eller optimering af maskinlæringsmodeller, hvilket fører til diskriminerende resultater eller skæve forudsigelser.

Årsager til modeldrevet bias

- 1. Bias ved valg af funktioner:** Modelfunktioner, der er valgt under modelleringsprocessen, kan utilsigtet kode bias, der findes i træningsdataene, hvilket fører til forudindtagede forudsigelser eller diskriminerende resultater.
- 2. Algoritmisk kompleksitet:** Komplekse maskinlæringsalgoritmer kan opfange og forstærke subtile skævheder i træningsdataene og forstærke deres indvirkning på modellens forudsigelser.
- 3. Optimeringsmålsætninger:** Optimeringsmål defineret under modeltræningsprocessen kan utilsigtet prioritere visse resultater frem for andre, hvilket fører til forudindtagede eller uretfærdige forudsigelser.





## Eksempler på modeldrevet bias

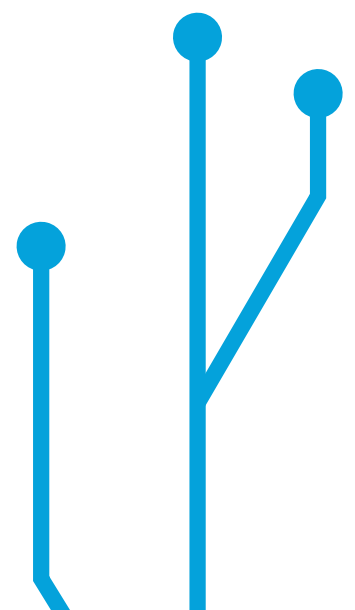
- 1. Kønsbias i ansættelsesalgoritmer:** Automatiserede ansættelsesalgoritmer kan utilsigtet favorisere mandlige kandidater frem for kvindelige kandidater på grund af forudindtaget funktionsvalg eller optimeringsmål, hvilket opretholder kønsforskelle i arbejdsstyrken.
- 2. Racemæssig bias i domsalgoritmer:** Prædiktive domsalgoritmer, der bruges i strafferetlige systemer, kan uforholdsmæssigt anbefale hårdere straffe til minoritetstiltalte og forstærke racemæssige forskelle i fængslingsprocenter.
- 3. Socioøkonomisk bias i modeller for lånegodkendelse:** Maskinlæringsmodeller, der bruges til lånegodkendelse, kan systematisk afvise lån til personer fra marginaliserede samfund, hvilket forværrer de socioøkonomiske uligheder i adgangen til finansielle tjenester.



## Indvirkning af modeldrevet bias

- 1. Fortsættelse af diskrimination:** Modeldrevet bias kan fastholde og forstærke eksisterende diskrimination og uligheder i samfundet, hvilket fører til uretfærdig behandling og forudindtagede resultater for marginaliserede grupper.
- 2. Mangel på ansvarlighed:** Fordomsfulde AI-modeller kan mangle gennemsigtighed og ansvarlighed, hvilket gør det udfordrende at identificere og håndtere diskriminerende praksis i AI-systemer.
- 3. Ethiske konsekvenser:** Modeldrevet bias rejser etiske spørgsmål om fairness, retfærdighed og menneskerettigheder, hvilket understreger behovet for etiske retningslinjer og regler for udvikling og anvendelse af kunstig intelligens.

Modeldrevet bias udgør en betydelig udfordring for udviklingen og implementeringen af retfærdige og ansvarlige AI-systemer. Ved at forstå mekanismerne og konsekvenserne af modeldrevet bias kan interessenter implementere strategier til at afbøde bias og fremme fairness og retfærdighed i AI-teknologier.





## > Menneskedrevet bias

### Hvad er menneskedrevet bias?

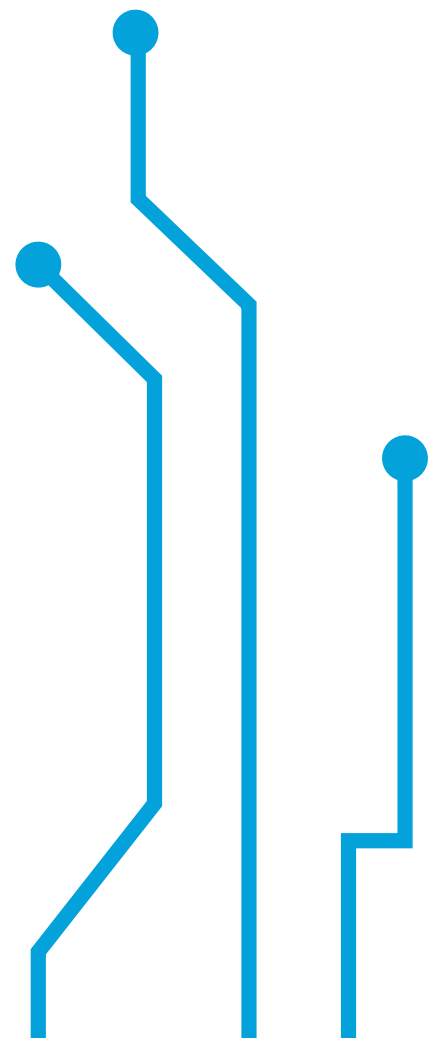
Menneskedrevet bias i AI refererer til bias, der opstår som følge af beslutninger, handlinger eller vurderinger fra personer, der er involveret i udvikling og implementering. Det kan stamme fra kognitive bias, kulturelle påvirkninger og samfundsmæssige fordomme, der fører til forudindtagede resultater eller diskriminerende praksis.

### Årsager til menneskedrevet bias

- 1. Skævheder i dataindsamlingen:** Skævheder i dataindsamlingen som f.eks. prøveudtagning eller udvælgelse kan føre til skæve træningsdata og skæve modelforudsigelser.
- 2. Bias i algoritmisk design:** AI-algoritmer kan være forudindtagede på grund af designere og udvikleres valg, hvilket viderefører forudindtagede resultater i AI-systemer.
- 3. Fortolknings- og implementeringsbias:** Menneskelige tolke og beslutningstagere kan være forudindtagede, når de anvender AI-systemer, hvilket kan føre til diskriminerende praksis og uretfærdig behandling.

## Eksempler på menneskedrevet bias

- 1. Bias i ansigtsgenkendelsessystemer:** Menneskelige skævheder i indsamlingen af træningsdata og det algoritmiske design kan føre til racemæssige eller køns-mæssige skævheder i ansigtsgenkendelsessystemer, hvilket resulterer i forkert identifikation eller underrepræsentation af visse demografiske grupper.
- 2. Retfærdighed i ansættelsesalgoritmer:** Skævheder i menneskelige beslutningsprocesser, som f.eks. screening af CV'er eller evaluering af interviews, kan fastholde køns- eller raceforskelle i ansættelsesresultater, selv når man bruger AI-baserede ansættelsesalgoritmer.

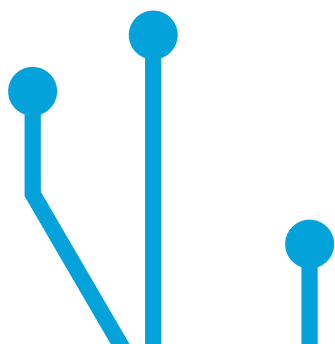




## Virkningen af menneskeskabte fordomme

- 1. Forværring af eksisterende uligheder:** Menneskedrevne bias i AI kan forværre eksisterende uligheder og forskelle i samfundet. Partisk dataindsamling, algoritmisk design og fortolkning kan føre til uretfærdig behandling af marginaliserede grupper, fastholde diskrimination og hindre sociale fremskridt.
- 2. Erosion af tillid og offentlighedens tiltro:** AI-systemer, der er præget af menneskelige fordomme, kan undergrave offentlighedens tillid til teknologien. Bekymringer om retfærdighed, gennemsigtighed og ansvarlighed kan opstå og hindre indførelsen og accepten af AI i forskellige sektorer.
- 3. Reduceret effektivitet af AI-systemer:** Menneskedrevne fordomme kan underminere AI-systemers effektivitet. Biased træningsdata eller biased fortolkninger af mennesker kan føre til unøjagtige forudsigelser, fejlbehæftede anbefalinger og suboptimale resultater, hvilket forhindrer de potentielle fordele ved AI.

Menneskelig bias er en væsentlig udfordring for at skabe retfærdige og ansvarlige AI-systemer. Ved at forstå og afbøde algoritmiske bias kan interessenter opbygge mere troværdige og gennemsigtige AI-systemer.





# Charlæ



Universitat  
de les Illes Balears



ISQe  
ENGAGING PEOPLE



INNOVATION TRAINING CENTER



AARHUS UNIVERSITY



VAMK  
VAASAN AMMATTIKORKEAKOULU  
UNIVERSITY OF APPLIED SCIENCES



helixconnect



Medfinansieret af  
Den Europæiske

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



2022-1-ES01-KA220-HED-000085257