



# Etisk AI-mikrocertifikat

HÆFTE

CU2 | Non-maleficence

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

# 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**

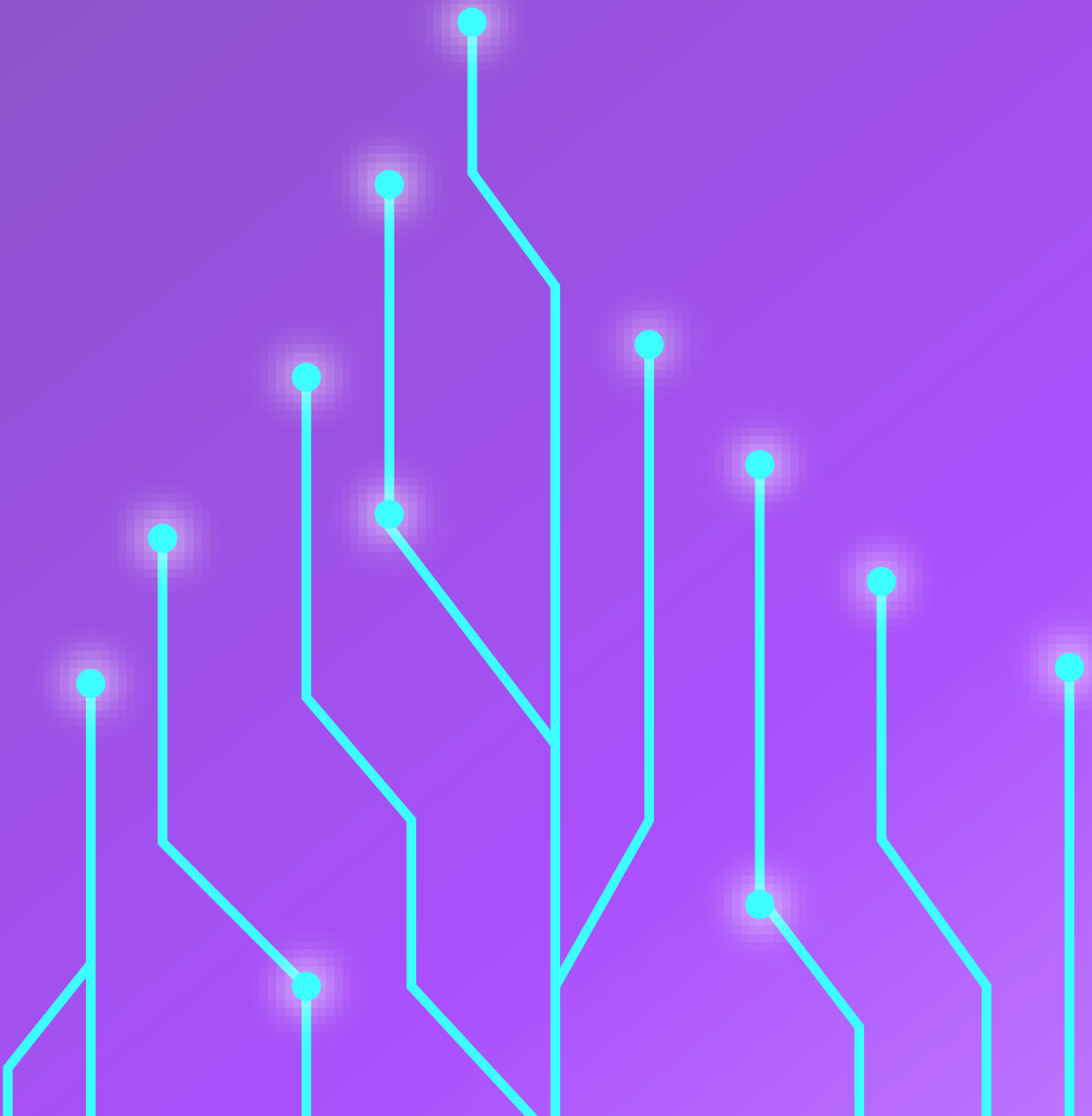
**02. Non-maleficence**

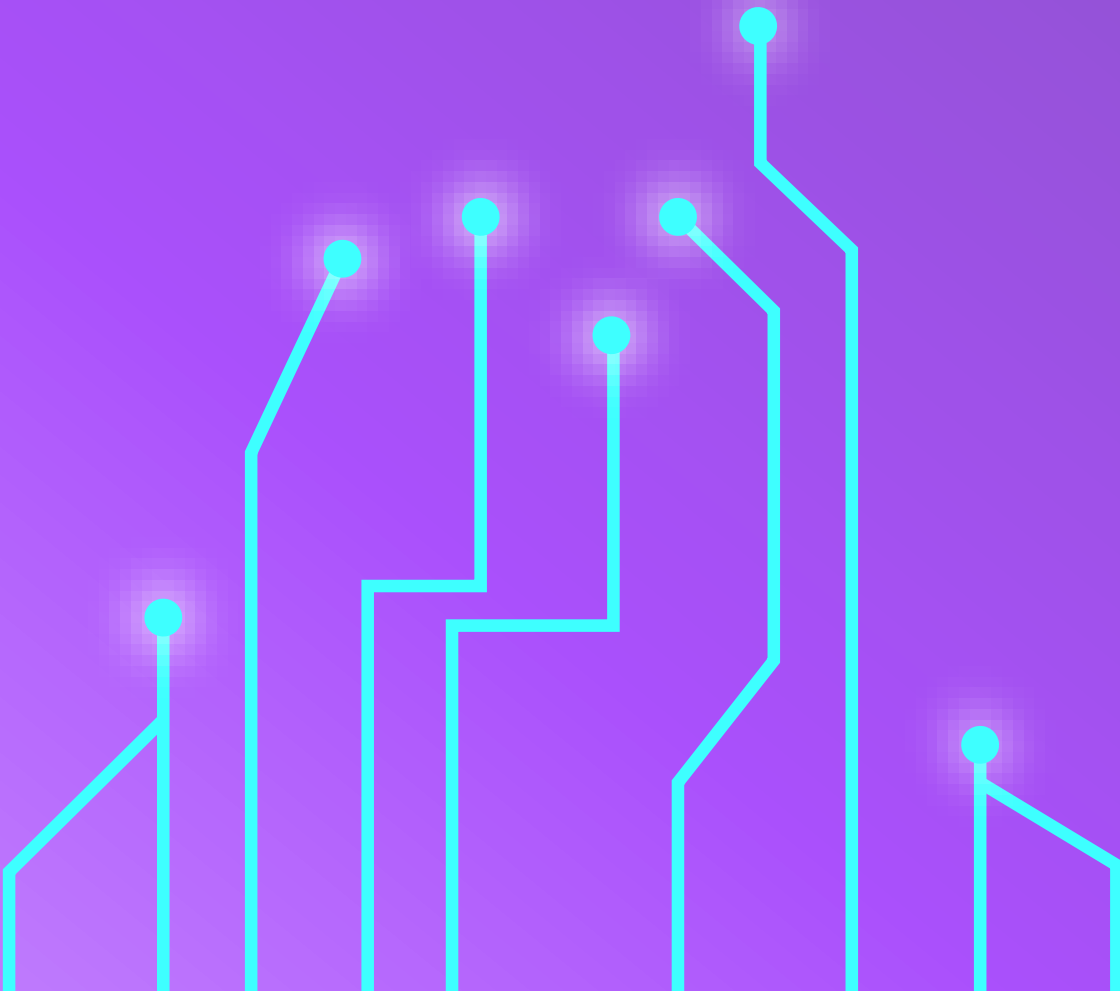
**03. Mulige skader fra forudindtaget AI**

**04. Strategier til at gøre AI-systemer mindre skadelige**

# 01. Introduction

CU2 | Non-maleficence



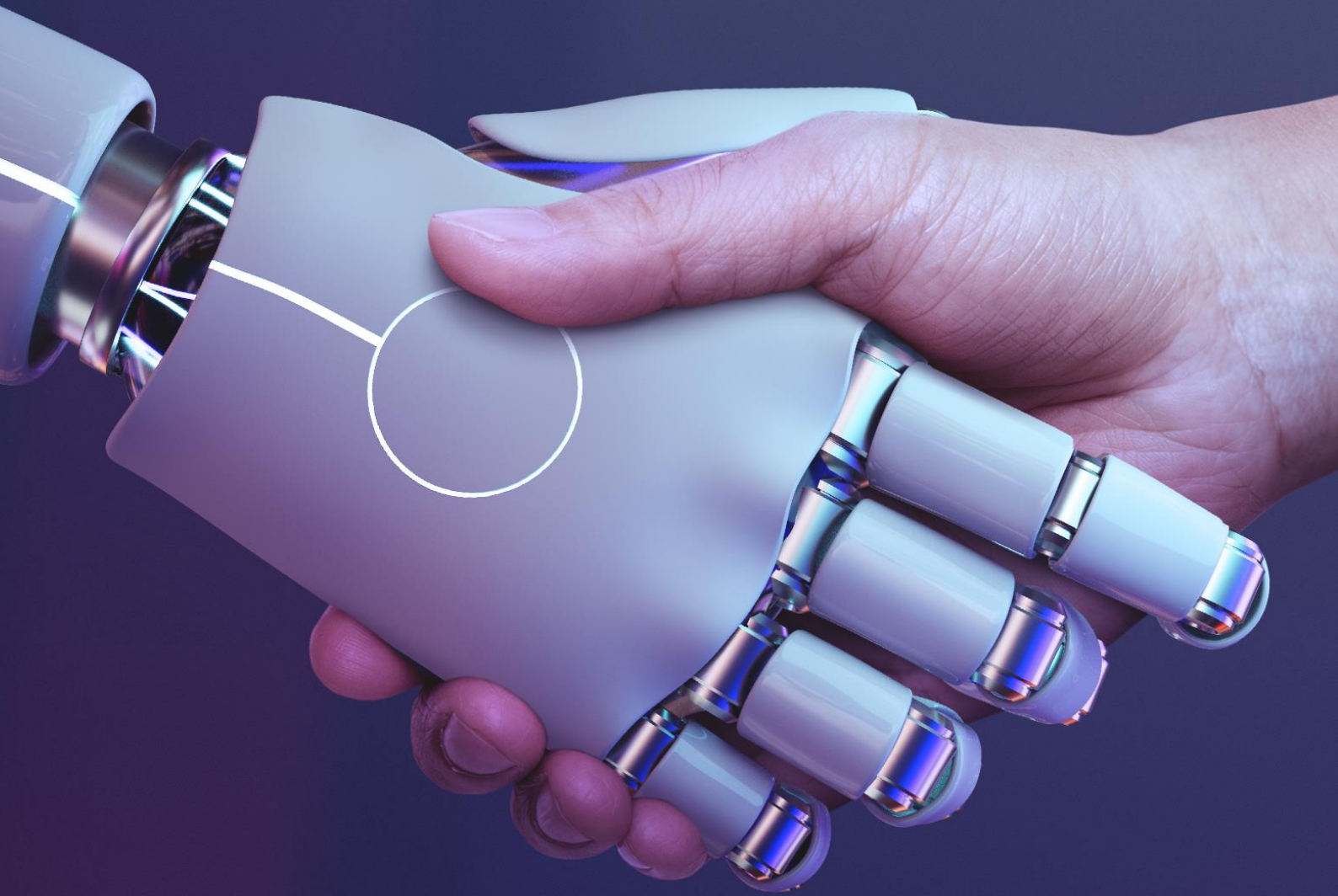


## 01. Introduktion

I denne kompetenceenhed vil eleverne få grundlæggende viden om begrebet non-maleficence i AI, AI-udviklers og -brugers ansvar for at sikre etiske AI-systemer med minimal skade og anerkende den virkelige verdens konsekvenser ved at værdsætte vedtagelsen og implementeringen af mekanismer, der fremmer ansvarlighed i AI-systemer.

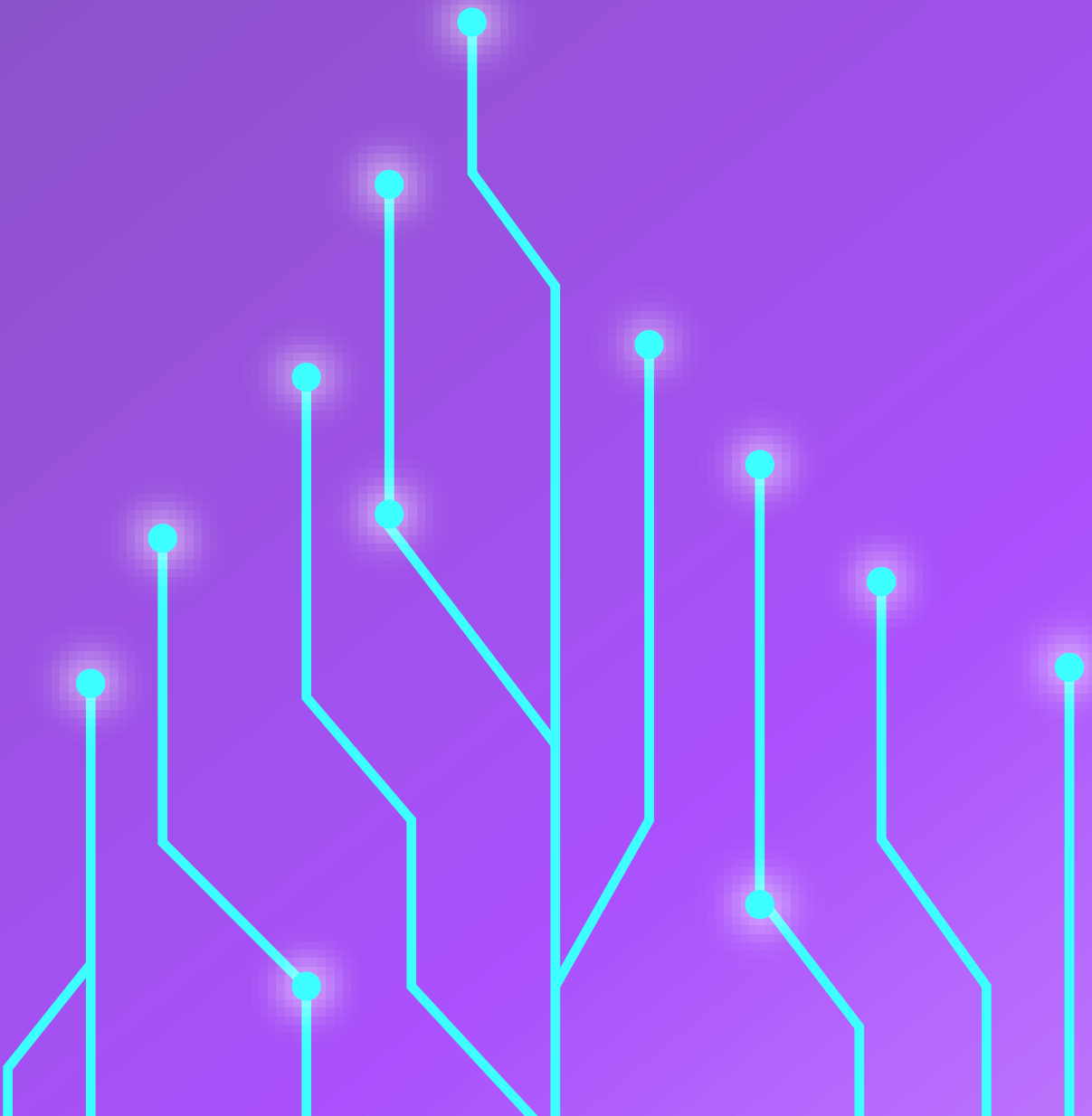
Vidensmålene for denne kompetenceenhed omfatter:

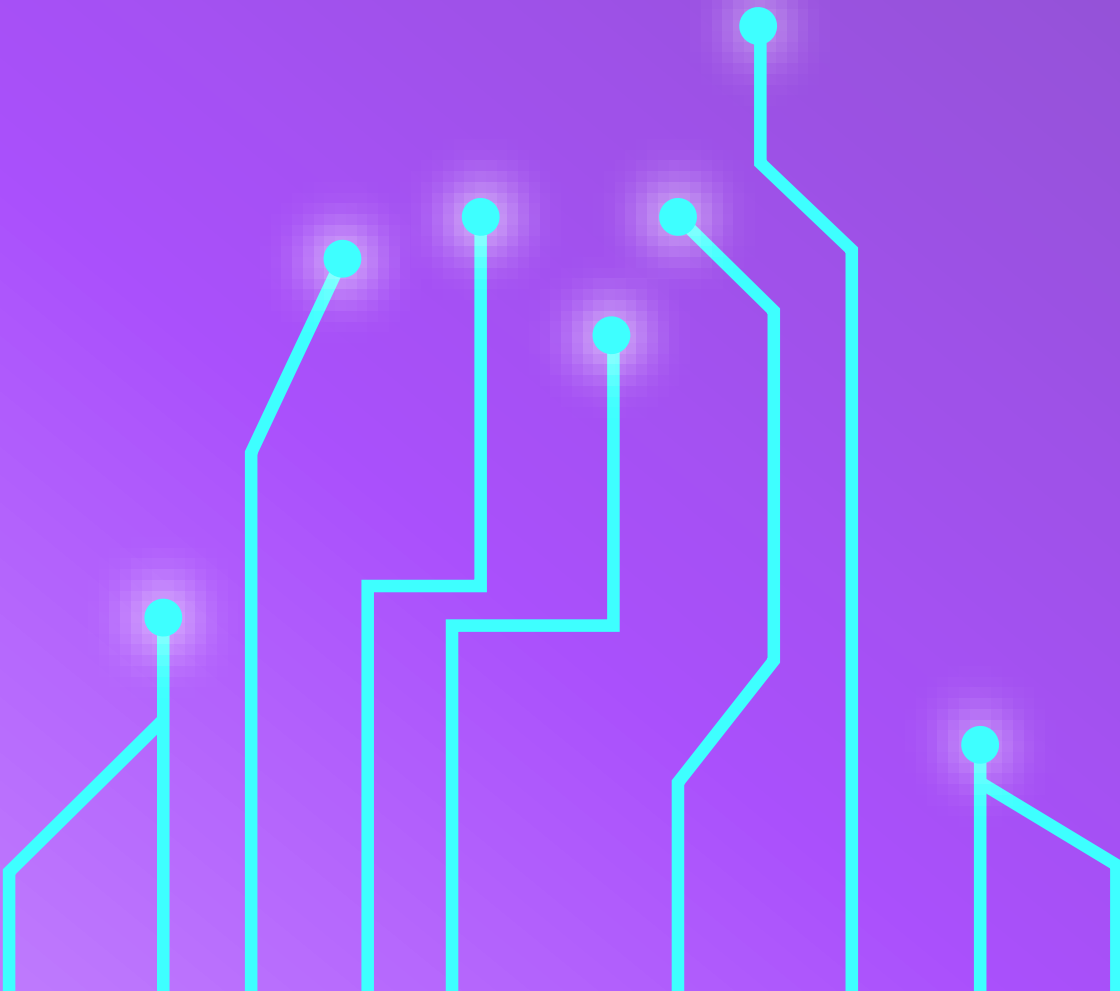
- **Non-maleficence-princippet:** Eleverne lærer det grundlæggende non-maleficence-princip, der understreger vigtigheden af at undgå skade, når man skaber og bruger AI-systemer, og hvordan denne idé bidrager til ansvarlig AI-udvikling.
- **Mulige skader fra forudindtaget AI:** Eleverne vil genkende de forskellige måder, hvorpå forudindtagede AI-systemer kan forårsage skade, såsom diskrimination eller krænkelse af privatlivets fred, og bruge eksempler fra den virkelige verden til at illustrere betydningen af at håndtere algoritmisk forudindtagethed.
- **Strategier til at gøre AI-systemer mindre skadelige:** Eleverne vil blive fortrolige med enkle strategier, der kan gøre AI-systemer mindre skadelige, herunder fremme af retfærdighed, ansvar og gennemsigtighed i AI-udvikling og tilskyndelse til samarbejde med eksperter fra forskellige områder.



# 02. Non-maleficence

CU2 | Non-maleficence





## 02. Non-maleficence

I dette afsnit introducerer vi non-maleficence-princippet og dets relevans for AI og big data-teknologier. Princippet om non-maleficence, der ofte opsummeres som "gør ingen skade", er en hjørnesten i etisk beslutningstagning inden for forskellige områder, herunder medicin, teknologi og forskning. I forbindelse med AI og big data understreger non-maleficence vigtigheden af at prioritere individets og samfundets sikkerhed og velbefindende, når disse teknologier udvikles og implementeres.

### > Hvad er non-maleficence?

Non-maleficence, der stammer fra den latinske sætning "primum non nocere", som betyder "først, gør ingen skade", er et grundlæggende etisk princip, der vejleder fagfolk i at forhindre skade på andre. Det understreger den moralske forpligtelse til at undgå at forårsage skade, hvad enten det er fysisk, psykologisk eller samfundsmæssigt, gennem ens handlinger eller beslutninger. I forbindelse med AI og big data kræver non-maleficence, at udviklere, forskere og politiske beslutningstagere overvejer de potentielle risici og konsekvenser af AI-teknologier og træffer proaktive foranstaltninger for at forhindre skade.





## > **Hvorfor er non-maleficence vigtigt?**

Non-maleficence er særligt vigtigt inden for AI og big data på grund af den betydelige indvirkning, disse teknologier kan have på individer og samfund. AI-systemer bruges i stigende grad i kritiske beslutningsprocesser som f.eks. sundhedsdiagnoser, finansielle udlån og strafferetlige domme. At sikre, at disse systemer prioriterer etiske overvejelser og ikke forårsager skade, er afgørende for at opretholde offentlighedens tillid, forhindre diskrimination og opretholde samfundsmæssige værdier som retfærdighed og fairness.

## > **Principper for ikke-fejlbarlighed**

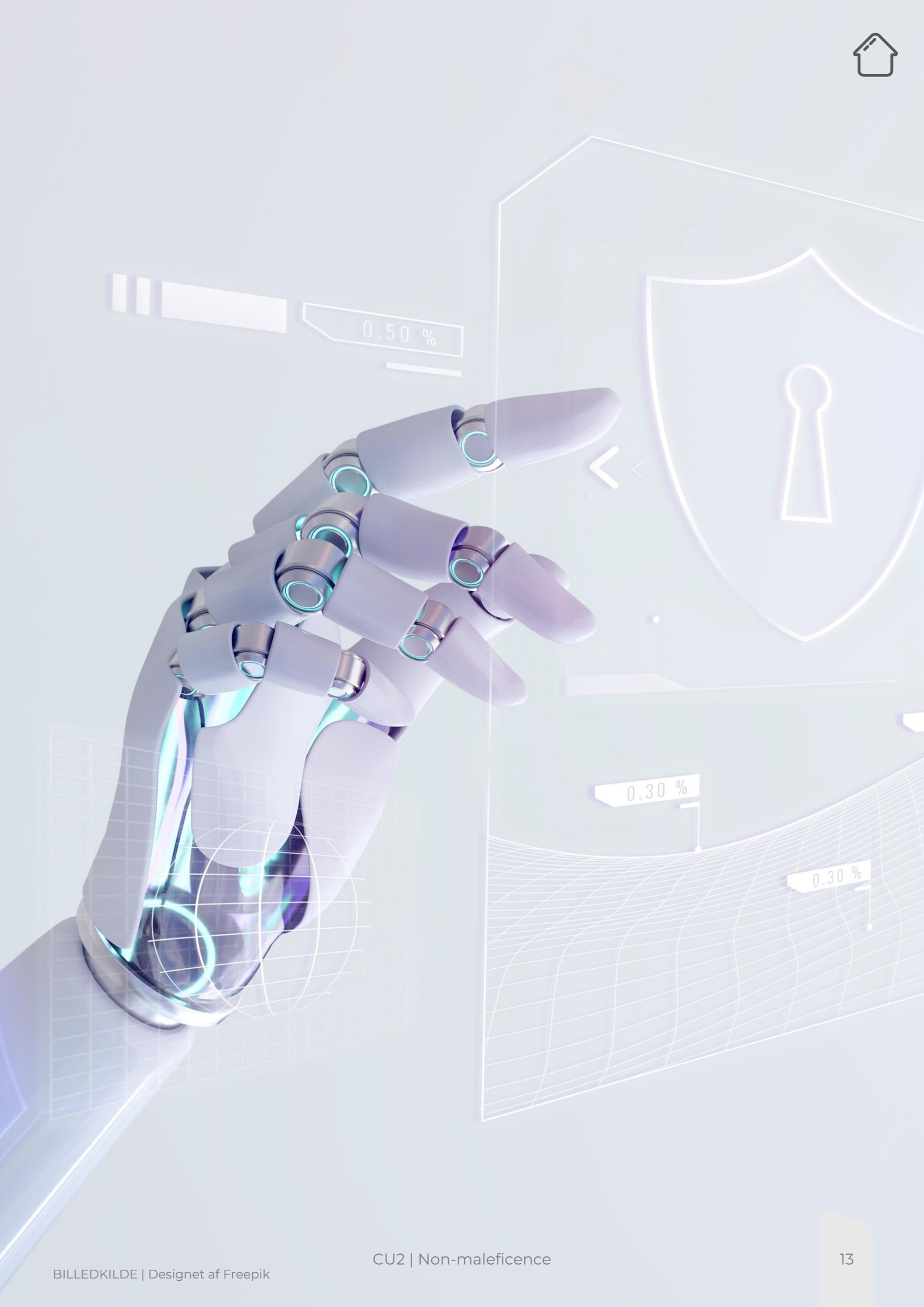
Non-maleficence kræver, at personer og organisationer, der er involveret i AI-udvikling, aktivt identificerer og afbøder potentielle skader, som AI-systemer kan udgøre. Det indebærer, at man ikke kun overvejer de umiddelbare konsekvenser af AI-teknologier, men også deres langsigtede konsekvenser og utilsigtede virkninger. Non-maleficence tilskynder til en proaktiv tilgang til etik, hvor udviklere forudser og håndterer potentielle risici, før de materialiserer sig.



## > **Anvendelse i AI-udvikling**

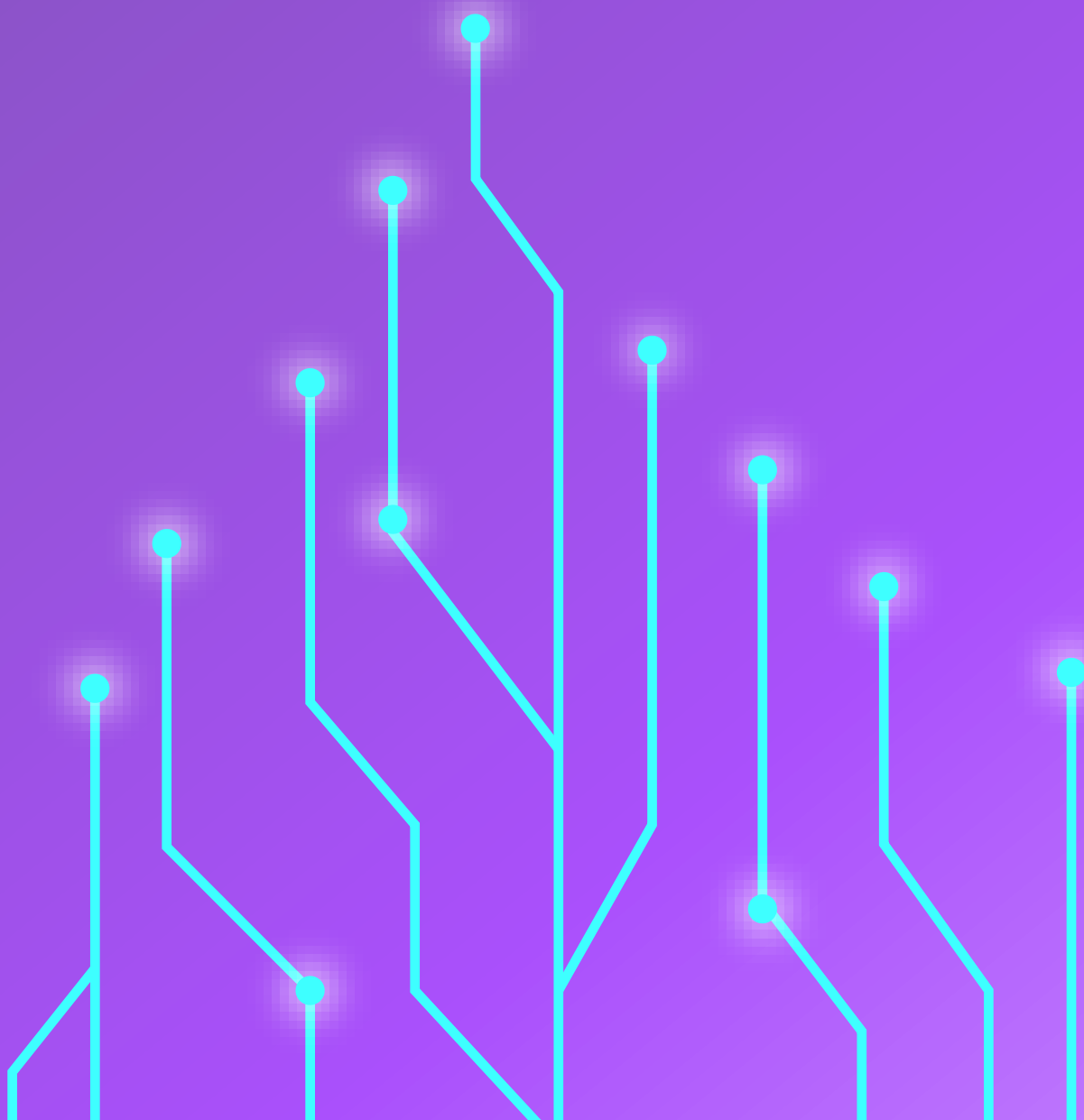
I forbindelse med AI-udvikling kommer non-maleficence til udtryk gennem forskellige praksisser, der har til formål at minimere skade og fremme etisk brug. Det omfatter strenge test- og valideringsprocedurer for at identificere og korrigere bias i AI-algoritmer, gennemsigtig dokumentation af AI-systemers beslutningsprocesser for at øge ansvarligheden og løbende overvågning og evaluering af AI-implementeringer for at sikre, at de er i overensstemmelse med etiske standarder og samfundsmæssige værdier.

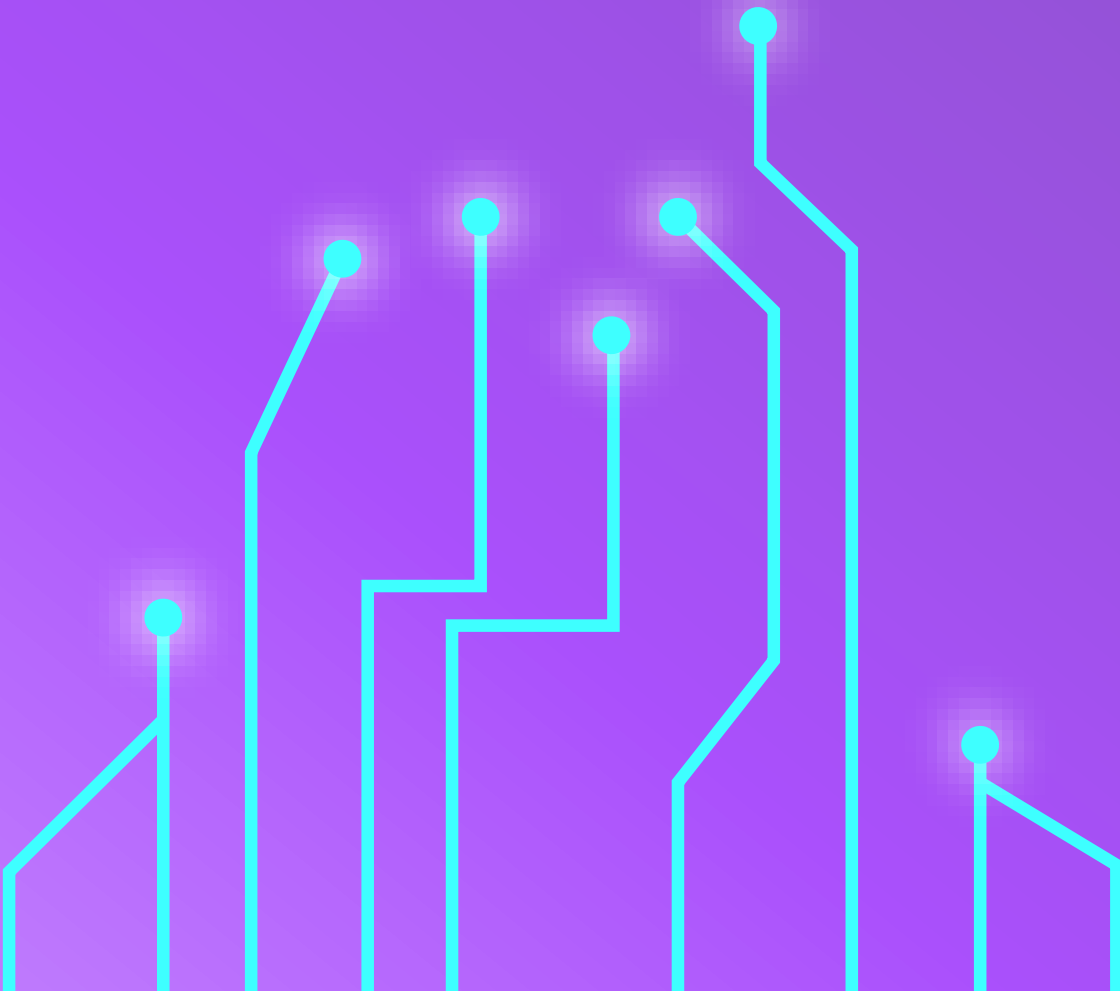




# 03. Mulige skader fra forudindtaget AI

CU2 | Non-maleficence





### 03. Mulige skader fra forudindtaget AI

I dette afsnit vil vi udforske de forskellige måder, hvorpå forudindtagede AI-systemer kan forårsage skade, lige fra diskrimination til krænkelse af privatlivet. At forstå disse potentielle skader er afgørende for at anerkende vigtigheden af at håndtere algoritmisk bias og fremme ansvarlig AI-udviklingspraksis.

#### > **Anerkendelse af skadelige virkninger**

Biased AI-systemer har potentiale til at fastholde og forværre eksisterende uligheder og uretfærdigheder i samfundet. Forestil dig en verden, hvor en algoritme uretfærdigt nægter dig et lån på grund af dit postnummer, eller hvor et ansigtsgenkendelsessystem fejlidentificerer dig som kriminel på grund af racefordomme. Det er blot nogle få af de potentielle farer, som partisk AI udgør. Nedenfor udforsker vi ti af de mest almindelige skadelige scenarier, der kan opstå som følge af forudindtagede AI-systemer.

- 1. Diskriminerende resultater:** Biased AI-algoritmer kan føre til diskriminerende resultater, hvor visse individer eller grupper behandles uretfærdigt baseret på karakteristika som race, køn eller socioøkonomisk status. Dette kan resultere i forskelle på forskellige områder, herunder beskæftigelse, uddannelse og strafferet.

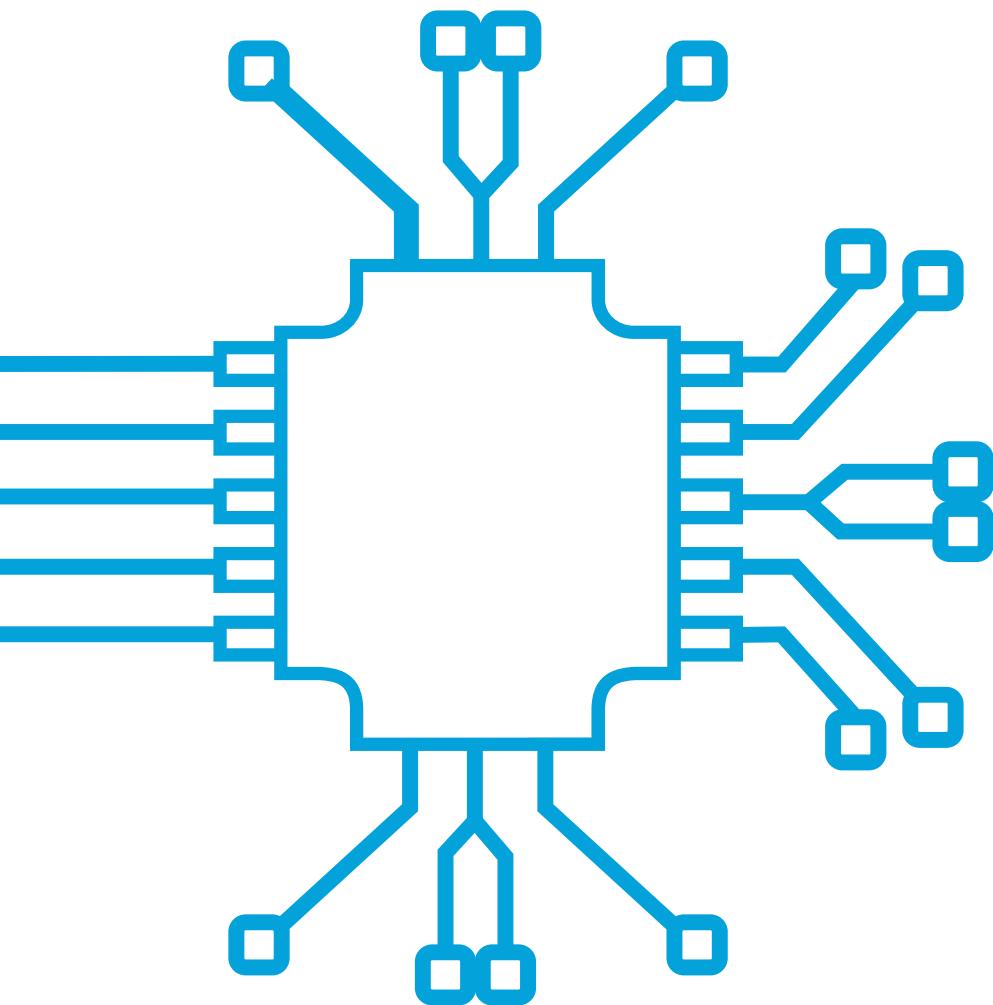


- 2. Krænkelser af privatlivets fred:** Biased AI-systemer kan krænke enkeltpersoners ret til privatliv ved at træffe beslutninger baseret på følsomme persondata uden deres samtykke. For eksempel kan ansigtsgenkendelsesteknologi, der anvendes i det offentlige rum, udsætte enkeltpersoner for uberettiget overvågning og sporing, hvilket giver anledning til bekymring for krænkelser af privatlivet og borgerlige frihedsrettigheder.
- 3. Forstærkning af stereotyper:** Biased AI-algoritmer kan videreføre og forstærke skadelige stereotyper og fordomme i samfundet. Det kan føre til marginalisering og stigmatisering af visse grupper, forværre eksisterende uligheder og hæmme sociale fremskridt.
- 4. Unøjagtig beslutningstagning:** Skævheder i træningsdata eller fejlbehæftede algoritmer kan resultere i unøjagtige eller fejlagtige beslutninger fra AI-systemer. Det kan have alvorlige konsekvenser, især inden for kritiske områder som sundhedsdiagnoser, finansiel rådgivning og strafferetlige domme, hvor forkerte beslutninger kan skade enkeltpersoner og samfundet.
- 5. Mangel på ansvarlighed:** Biased AI-systemer kan mangle gennemsigtighed og ansvarlighedsmekanismer, hvilket gør det vanskeligt at identificere og rette op på tilfælde af bias. Det kan underminere tilliden til AI-teknologier og hindre bestræbelserne på at håndtere algoritmisk bias effektivt.

- 6. Begrænset mangfoldighed og inklusion:** Biased AI-algoritmer kan fastholde eksisterende uligheder ved at favorisere visse demografiske grupper frem for andre. Det kan bidrage til manglende mangfoldighed og inklusion i udviklingen og implementeringen af AI, hvilket begrænser den repræsentation og de perspektiver, der afspejles i AI-systemer, og forværrer sociale uligheder.
- 7. Negativ indvirkning på innovation:** Fordomsfulde AI-algoritmer kan hindre innovation og fremskridt ved at fastholde forældede eller diskriminerende praksisser og begrænse mulighederne for kreativitet og udforskning. Det er vigtigt at tackle bias i AI for at skabe et miljø, der tilskynder til mangfoldighed i tankegangen og fremmer innovation til gavn for samfundet som helhed.
- 8. Tab af tillid og tiltro:** Tilfælde af bias i AI-systemer kan underminere offentlighedens tillid til teknologien og dens evne til at tjene det fælles bedste. Det kan føre til skepsis, modstand og modvilje mod at tage AI-løsninger i brug, hvilket hindrer deres potentiale til at påvirke samfundet positivt.
- 9. Juridiske og etiske bekymringer:** Biased AI-systemer kan give anledning til juridiske og etiske bekymringer i forhold til retfærdighed, ansvarlighed og gennemsigtighed. For at løse disse problemer er det nødvendigt med robuste lovgivningsmæssige rammer, etiske retningslinjer og ansvarlig AI-udviklingspraksis for at sikre, at AI-teknologier er i overensstemmelse med samfundets værdier og respekterer grundlæggende rettigheder.



**10. Sociale og økonomiske konsekvenser:** Den gennemgribende virkning af partisk AI strækker sig ud over individuelle tilfælde af diskrimination til bredere sociale og økonomiske konsekvenser. Fordomsfulde AI-systemer kan forværre eksisterende uligheder, udvide den digitale kløft og fastholde sociale uretfærdigheder, hvilket giver betydelige udfordringer for opbygningen af et fair og retfærdigt samfund.



## > **Eksempler fra den virkelige verden**

Ved hjælp af eksempler fra den virkelige verden vil vi illustrere betydningen af at adressere algoritmisk bias og dens potentielle indvirkning på enkeltpersoner og samfundet. Disse eksempler vil fremhæve tilfælde, hvor forudindtagede AI-systemer har ført til skadelige konsekvenser, såsom uretmæssige anholdelser, uretfærdig behandling i ansættelses- eller udlånsbeslutninger og fastholdelse af stereotyper og fordomme.

- **EKSEMPEL #1 - Amazons algoritme diskriminerede kvinder**

Amazons AI-ansættelsesværktøj havde til formål at finde de bedste tech-talenter, men det endte med at filtrere kvinder fra. Hvorfor det? Algoritmen, der var trænet på tidligere CV'er (for det meste fra mænd), favoriserede nøgleord, der blev brugt af mænd, og straffede dem, der blev forbundet med kvinder. Dette fremhæver en stor AI-udfordring: Biased data fører til biased algoritmer. Ligesom en elev, der er afhængig af fejlbehæftede lærebøger, arver AI skævhederne fra sine træningsdata. Læs mere i: <https://www.reuters.com/article/idUSKCN1MK0AG/>





- **EKSEMPEL 2 - Algoritmisk racebias i forudsigelse af tilbagefaldsrate for kriminelle**

Forestil dig et værktøj, der kan forudsige, hvem der begår forbrydelser. I USA gør COMPAS netop det, men med et racemæssigt twist. Undersøgelser viser, at sorte tiltalte langt oftere betegnes som højrisiko end hvide tiltalte med samme baggrund. Hvorfor denne skævhed? COMPAS afspejler de samfundsmæssige uligheder, der allerede findes i anholdelsesdata. Denne skævhed fører til, at folk bliver tilbageholdt før retssagen eller får hårdere domme, hvilket påvirker sorte personer uretfærdigt. Sagen om COMPAS understreger behovet for omhyggelig kontrol af AI, der bruges i retssystemer, for at sikre retfærdighed for alle. Læs mere på: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

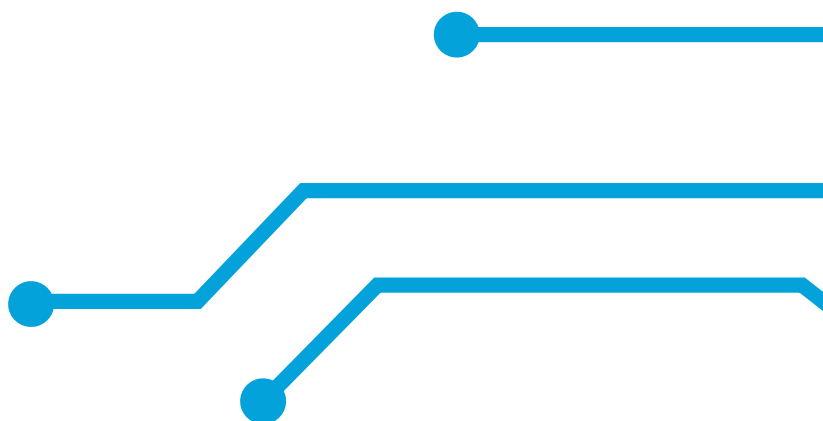


- **EKSEMPEL #3 - Amerikansk sundhedsalgoritme undervurderede sorte patienters behov**

Tænk på et sundhedssystem, der favoriserer patienter, som betaler mere. Desværre påvirkede det sorte patienter i USA. En algoritme, der er designet til at identificere dem, der har brug for ekstra pleje, overså mange sorte patienter på grund af bias. Hvorfor? Systemet baserede sig på tidligere data om medicinudgifter, som ikke afspejler sorte patienters begrænsede adgang til forebyggende behandling på grund af økonomiske uligheder. Det resulterede i, at sorte patienter blev kategoriseret som sundere og gik glip af kritisk pleje. En rettelse af algoritmen kunne hjælpe mange flere sorte patienter. Denne sag understreger behovet for fair AI i sundhedsvæsenet for at sikre, at alle får den behandling, de har brug for.

Læs mere på:

<https://www.theguardian.com/society/2019/oct/25/healthcare-algorithm-racial-biases-optum>





- **EKSEMPEL #4 - ChatBot delte diskriminerende beskeder**

Microsofts chatbot Tay var designet til at lære af almindelige samtaler. Da den blev lanceret på Twitter, begyndte den hurtigt at udspy racistiske og stødende beskeder. Hvorfor det? Fordi "trolls" bombarderede Tay med hadefuldt indhold, som den absorberede og efterlignede. Denne hændelse fremhæver en stor udfordring ved AI's interaktion med den virkelige verden. Sociale medier kan være et giftigt sted, og AI, der udsættes for det, kan have negativ læring. Tay er en advarende fortælling: Når man designer AI til online-interaktion, skal man overveje den sociale kontekst og potentialet for misbrug.

Læs mere i: <https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation>



- **EKSEMPEL #5 - Forudindtaget ansigtsgenkendelsessystem**

Forestil dig at fejre din fødselsdag med en shoppingtur og så blive anklaget for butikstyveri af et ansigtsgenkendelsessystem! Det skete for en Māori-kvinde i New Zealand. Teknologien, der er designet til at fange butikstyre, identificerede hende fejlagtigt og gav hende store problemer. Denne sag sætter fokus på farerne ved forudindtaget ansigtsgenkendelse. Undersøgelser viser, at disse systemer kan fejlidentificere mennesker, især kvinder og farvede. Efterhånden som ansigtsgenkendelsesteknologi bliver mere udbredt, er det afgørende at sikre retfærdighed og forhindre sådanne hændelser. Læs mere på:

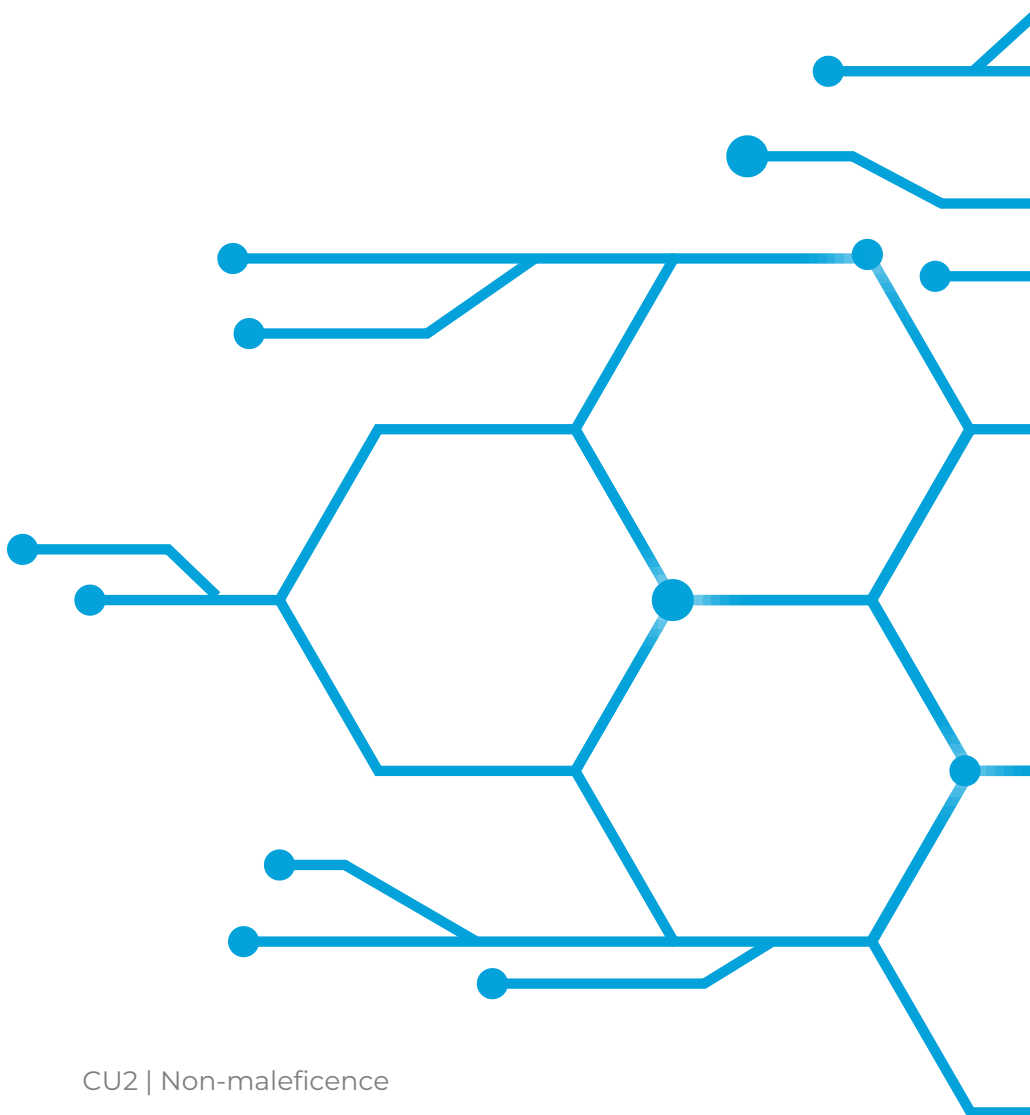
<https://www.1news.co.nz/2024/04/22/rotorua-mother-wrongly-identified-by-supermarket-as-a-thief/>



- **EKSEMPEL #6 - Generativ tekst AI opdigter fakta**

En juraprofessors omdømme blev plettet af en AI-chatbot. ChatGPT fabrikerede et krav om sexchikane mod ham med en falsk nyhedsartikel. Denne sag afslører en stor risiko ved AI: at generere skadelig misinformation. Professoren led skade på sit omdømme på trods af, at løgnen blev afsløret. Efterhånden som AI bliver mere almindeligt, er det vigtigt at sikre faktuelle oplysninger og fastlægge ansvaret for AI-genererede usandheder. Læs mere i:

<https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/>

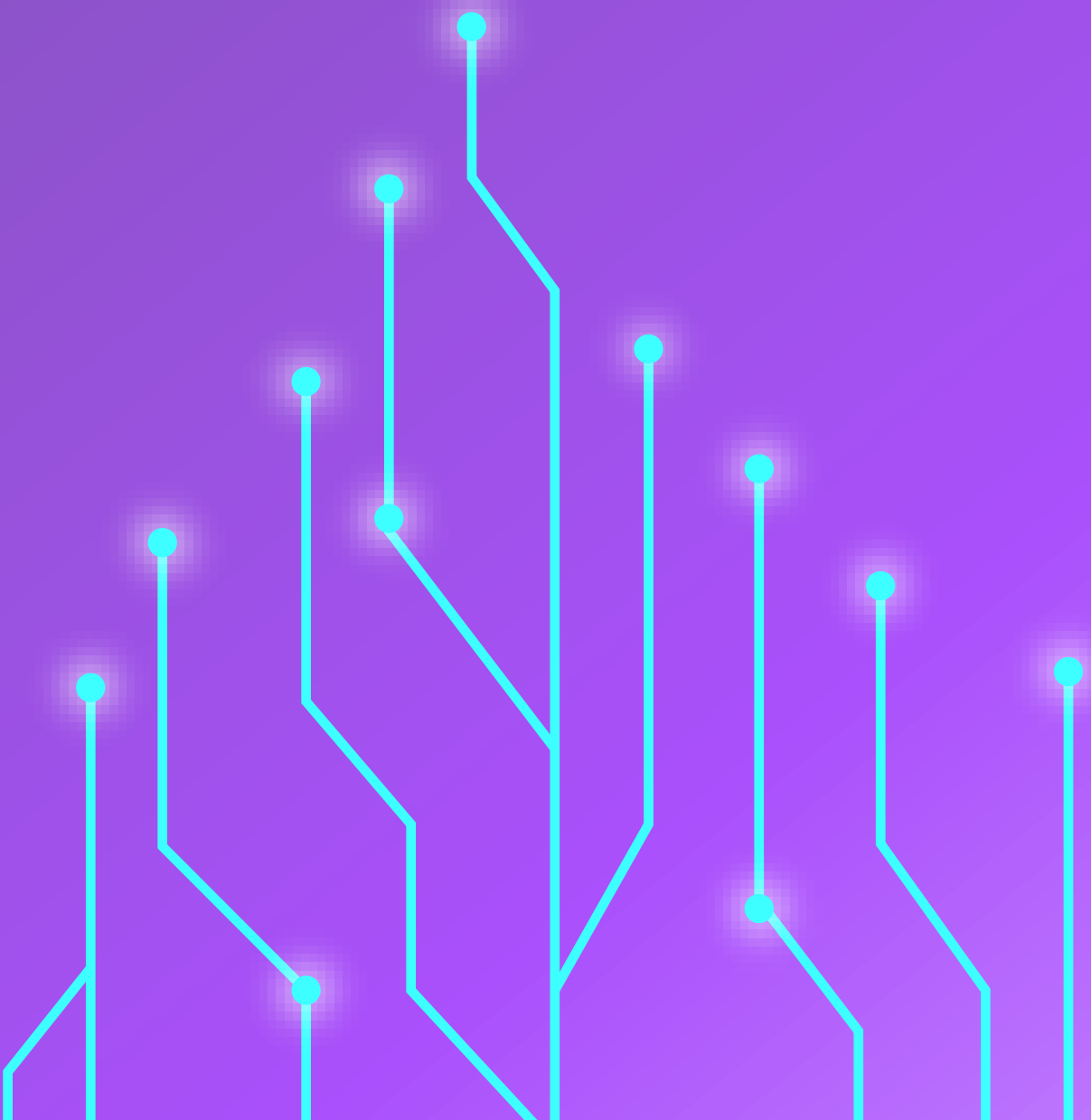


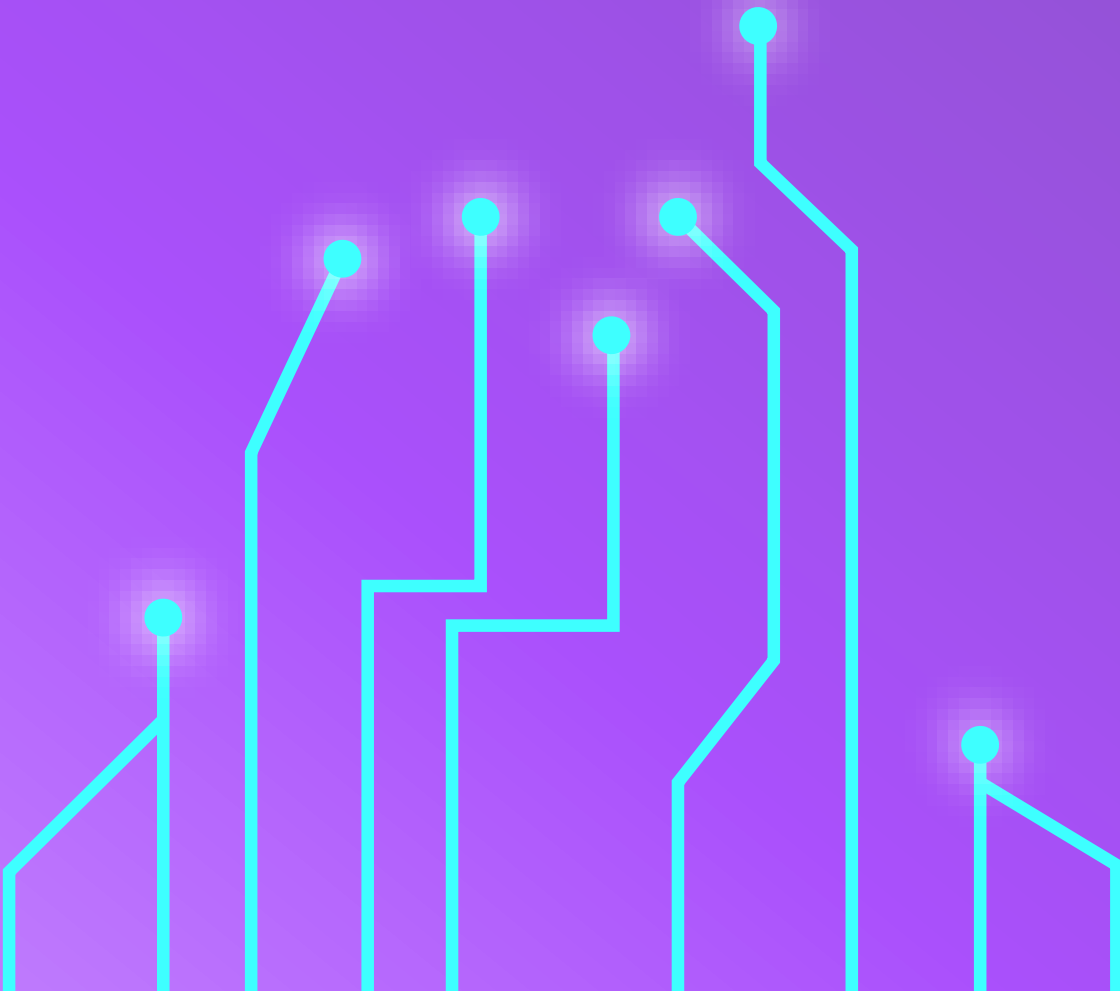


# AI

# 04. Strategier til at gøre AI-systemer mindre skadelige

CU2 | Non-maleficence





## 04. Strategier til at gøre AI-systemer mindre skadelige

I dette afsnit introducerer vi strategier, der har til formål at gøre AI-systemer mindre skadelige ved at fremme retfærdighed, ansvar og gennemsigtighed i deres udvikling og anvendelse. Disse strategier giver udviklere, politiske beslutningstagere og interessenter mulighed for proaktivt at håndtere algoritmisk bias og afbøde dens potentielle negative konsekvenser.

### > **Fremme af retfærdighed**

En vigtig strategi for at afbøde skader fra forudindtagede AI-systemer er at fremme retfærdighed i algoritmiske beslutningsprocesser. Det indebærer at sikre, at AI-modeller trænes på forskelligartede og repræsentative datasæt, der er fri for diskriminerende bias. Derudover kan retfærdigheds-bevidste maskinlæringsteknikker anvendes til at identificere og afbøde bias i algoritmiske forudsigelser og dermed fremme retfærdige resultater for alle individer.

### > **Øget ansvarlighed**

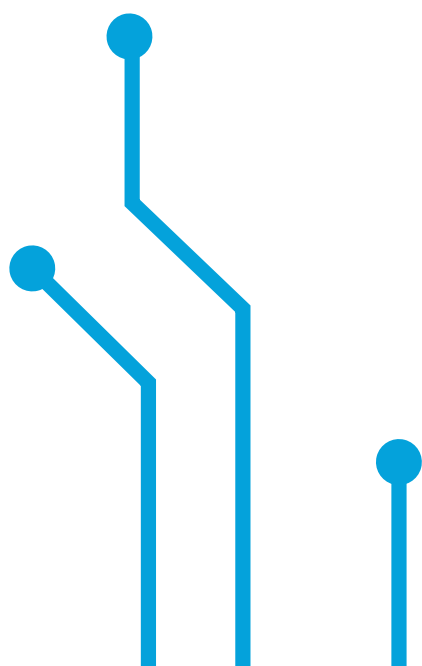
Et andet vigtigt aspekt af at reducere skader fra forudindtagede AI-systemer er at øge ansvaret blandt udviklere, organisationer og politiske beslutningstagere. Det omfatter implementering af etiske retningslinjer og den bedste praksis for AI-udvikling, f.eks. gennemførelse af grundige konsekvensanalyser for at identificere potentielle risici og skader.



Desuden kan etablering af klare ansvarlighedsmekanismer og tilsynsrammer hjælpe med at holde enkeltpersoner og organisationer ansvarlige for de etiske konsekvenser af deres AI-anvendelser. I den næste del af dette kursus vil vi udforske begrebet ansvarlighed mere detaljeret.

## > Tilskyndelse til gennemsigtighed

Gennemsigtighed er afgørende for at gøre AI-systemer mindre skadelige ved at fremme ansvarlighed og tillid blandt interessenter. Gennemsigtig dokumentation af AI-algoritmer og beslutningsprocesser giver mulighed for ekstern kontrol og validering, hvilket sikrer at skævheder og fejl identificeres og behandles rettidigt. Desuden kan fremme af åben dialog og samarbejde mellem AI-udviklere, forskere og berørte grupper lette større gennemsigtighed og forståelse af de etiske konsekvenser af AI-teknologier. Enhed 4 i dette kursus vil dykke dybere ned i begrebet gennemsigtighed, da det er et af de mest grundlæggende aspekter for at sikre ansvarlig AI.



## ➤ Sikring af privatlivet

AI-systemer er stærke værktøjer, men deres bekvemmelighed bør ikke komme på bekostning af privatlivet. Denne strategi fokuserer på at beskytte dine personlige oplysninger. Udviklere bør indsamle og bruge så få data som muligt, især følsomme oplysninger.

Sikkerhedsforanstaltningerne skal være i top for at holde oplysningerne sikre. AI-systemer skal også bygges, så de overholder love og regler om beskyttelse af personlige oplysninger, herunder den generelle forordning om databeskyttelse (GDPR) i Europa, som giver enkeltpersoner betydelig kontrol over deres personlige data.

## ➤ Prioritering af sikkerhed

Når det drejer sig om AI, bør sikkerhed have højeste prioritet. Det betyder, at AI-systemer skal gennemgå strenge test- og valideringsprocesser, før de slippes løs i den virkelige verden. Målet er at identificere og løse eventuelle potentielle risici eller problemer, der kan forårsage skade. Ved at sikre, at AI-systemer fungerer pålideligt og sikkert, kan vi beskytte enkeltpersoner og samfundet som helhed.







# Charlæ



Universitat  
de les Illes Balears



ISQe  
ENGAGING PEOPLE



itea  
INNOVATION TRAINING CENTER



AARHUS UNIVERSITY



VAMK  
VAASKA 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