



Etisk AI-mikrocertifikat

HÆFTE

CU4 | Gennemsigtighed

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. Vigtigheden af gennemsigtighed i AI-systemer

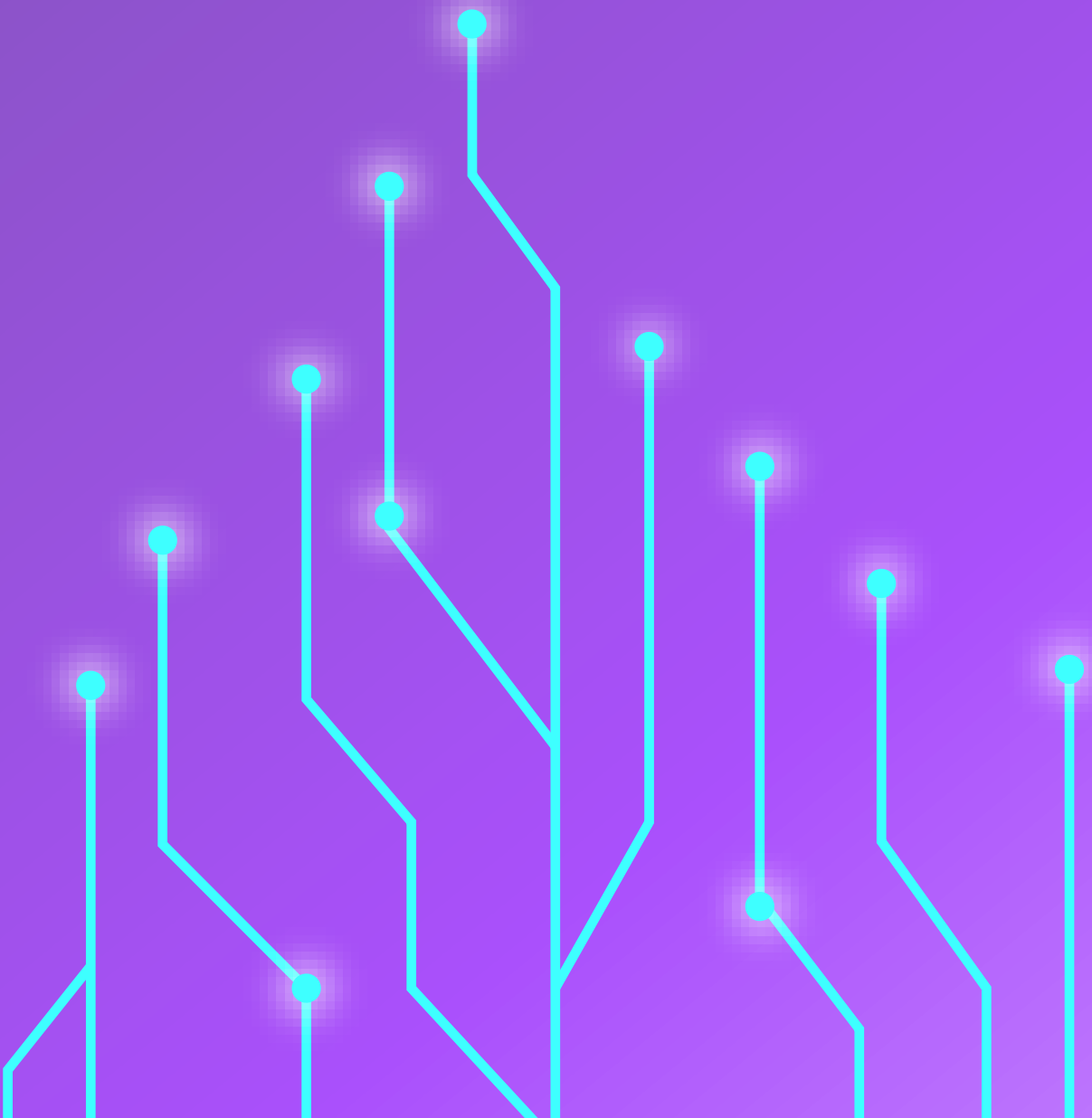
**03. Forholdet mellem gennemsigtighed og
algoritmisk bias**

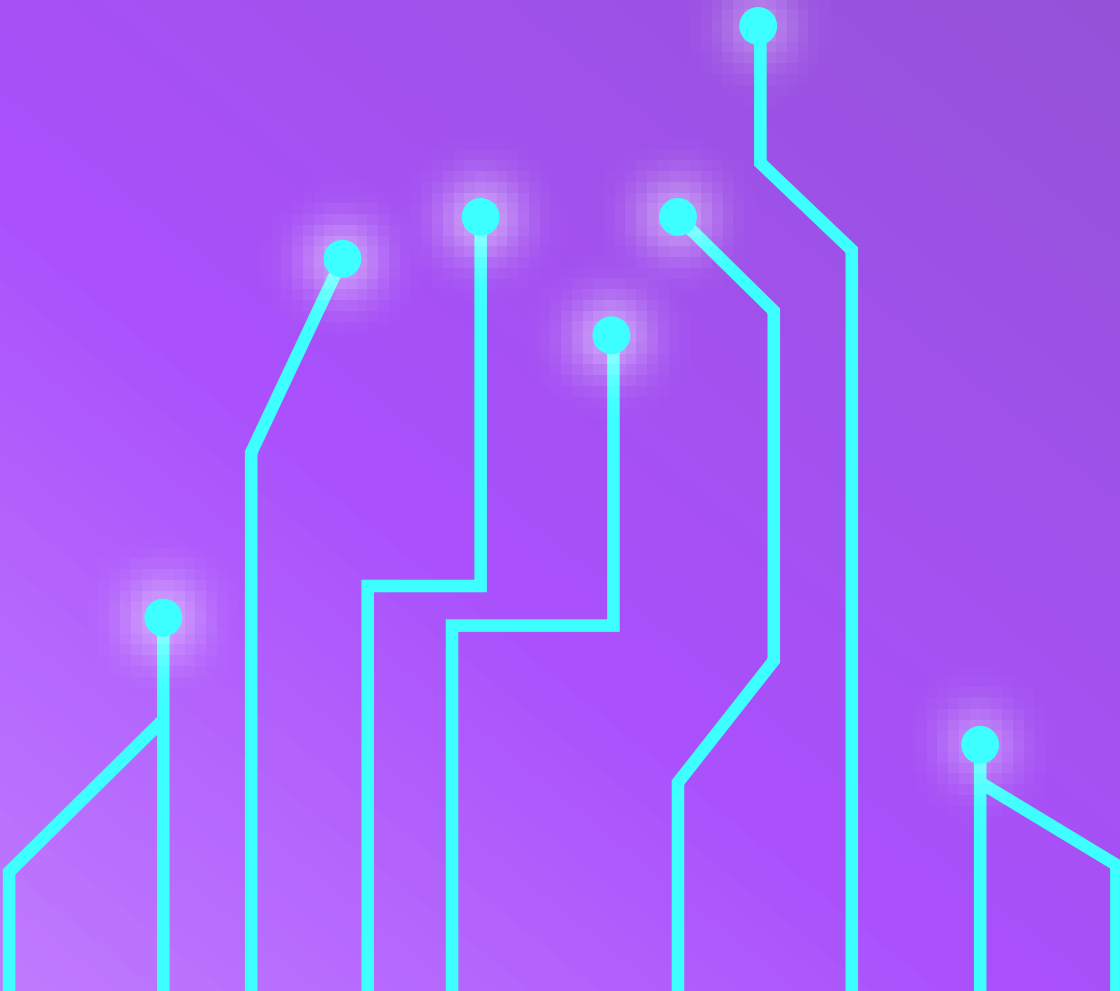
**04. Strategier til fremme af gennemsigtighed
i AI-systemer**

05. Konklusion

01. Introduktion

CU4 | Gennemsigtiged





01. Introduktion

I denne kompetenceenhed vil de studerende få viden om vigtigheden af gennemsigtighed i AI-systemer, med fokus på at forstå de grundlæggende begreber og forholdet mellem gennemsigtigheden og algoritmisk bias. Strategierne er vigtige at forstå for at sikre, at AI-systemer er forståelige, forklarlige og tilgængelige for interessenter, idet de anerkender konsekvenserne i den virkelige verden og værdsætter, hvor vigtigt fortolkelige modeller, klar dokumentation og effektiv kommunikation kan være for at fremme en kultur med gennemsigtighed og mindske algoritmisk bias.

Videnmålene for dette kursus omfatter:

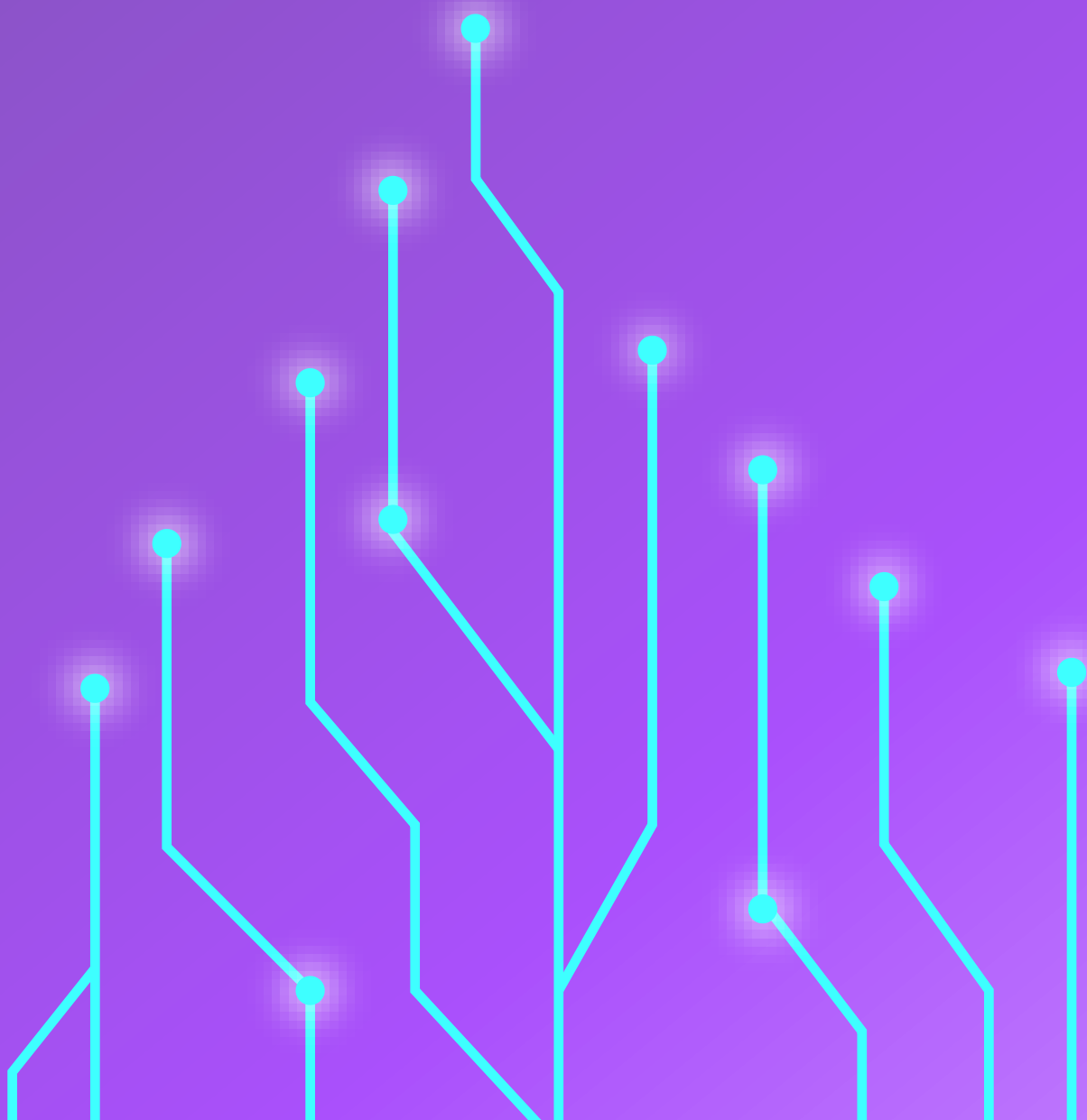
- **Betydningen af gennemsigtighed i AI-systemer** og dens relevans for at sikre, at AI-systemer er forståelige, forklarlige og tilgængelige for interessenter. Vi vil identificere fordelene ved og værdsætte betydningen af gennemsigtige AI-systemer for at opbygge tillid og muliggøre forståelse hos interessenter. Som et eksempel: En AI-model, der er designet til at opdage kræft, kan true et liv, selv om den kun tager 1 % fejl. I sådanne tilfælde er AI og mennesker nødt til at arbejde sammen, og opgaven bliver meget lettere, når AI-modellen kan forklare, hvordan den nåede frem til en bestemt beslutning. Gennemsigtighed gør AI til en holdspiller.



- **Forholdet mellem gennemsigtighed og algoritmisk bias** for at finde forbindelsen mellem gennemsigtighed og algoritmisk bias, anerkende farerne ved uigennemsigtighed, og hvordan øget gennemsigtighed kan hjælpe med at identificere, forhindre og afbøde forudindtagede resultater i AI-systemer. Vi vil anerkende betydningen af gennemsigtighed for at håndtere og afbøde algoritmisk bias. Som et eksempel: Ofte er AI-algoritmer uigennemsigtige i den forstand, at sådanne forklaringer ikke er tilgængelige for alle interessenter. Denne uigennemsigtighed kan have forskellige kilder. Nogle gange undlader institutioner eller virksomheder at kommunikere, når de er afhængige af AI-systemer, eller om hvordan disse systemer fungerer.
- **Strategier til fremme af gennemsigtighed i AI-systemer**, såsom brug af fortolkelige modeller, klar dokumentation og kommunikation af beslutningsprocesserne i AI-applikationer. Vi vil forklare, hvor vigtige disse strategier er for at fremme en kultur med gennemsigtighed og mindske algoritmisk bias. AI kan f.eks. påvirke forskellige interessenter som brugere, kunder, medarbejdere, ledere, tilsynsmyndigheder eller samfundet. For at sikre gennemsigtighed og ansvarlighed er du nødt til at engagere og styrke dine AI-interessenter gennem hele IS-livscyklussen.

02. Vigtigheden af gennemsigtighed i AI-systemer

CU4 | Gennemsigtighed





02. Vigtigheden af gennemsigtighed i AI-systemer

Gennemsigtighed er et grundlæggende princip i udviklingen og implementeringen af systemer med kunstig intelligens (AI).

Gennemsigtighed i AI henviser til åbenhed og tilgængelighed af AI-systemer, så interessenter kan forstå, hvordan algoritmer fungerer, hvorfor visse beslutninger træffes, og hvilke faktorer der påvirker deres output. Det omfatter forskellige aspekter, herunder tilgængeligheden af information om datakilder, algoritmiske modeller, beslutningsprocesser og potentielle bias. Gennemsigtige AI-systemer gør det muligt for interessenter, herunder brugere, udviklere, beslutningstagere og den brede offentlighed, at granske og udfordre algoritmiske resultater, hvilket fremmer tillid og ansvarlighed.

En af de vigtigste fordele ved gennemsigtige AI-systemer er deres forståelighed. Når AI-algoritmer er gennemsigtige, kan interessenter forstå, hvordan de fungerer, og hvorfor de giver bestemte resultater. Denne forståelse gør det muligt for brugerne at stole på AI-teknologier og træffe informerede beslutninger om deres brug. I forbindelse med en AI-model til medicinsk diagnose giver gennemsigtighed f.eks. sundhedspersonale mulighed for at forstå, hvordan modellen er nået frem til sin diagnose, så de kan validere dens nøjagtighed og pålidelighed, før de træffer beslutninger om behandling.



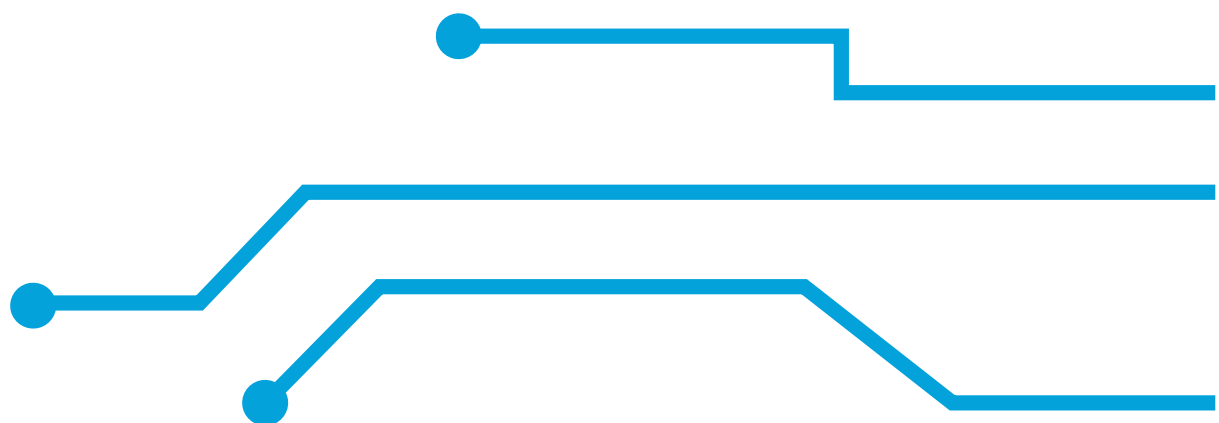


Desuden gør gennemsigtighed det lettere at forklare resultater, hvilket er afgørende for at sikre, at AI-systemer kan give fortolkelige forklaringer på deres beslutninger og handlinger. Forklarlig AI giver interessenter mulighed for at forstå rationalet bag algoritmiske resultater og for at identificere og korrigere bias eller fejl. Hvis der f.eks. er tale om et AI-system til godkendelse af lån, gør gennemsigtighed og forklarlighed det muligt for låneansøgere at forstå, hvorfor deres ansøgning blev godkendt eller afvist, hvilket giver indsigt i beslutningsprocessen og muligheder for at klage, hvis de mener, at beslutningen var forudindtaget eller uretfærdig.

Derudover forbedrer gennemsigtighed tilgængeligheden af AI-systemer, hvilket gør dem mere inkluderende og retfærdige. Når AI-algoritmer er gennemsigtige, kan interessenter med forskellige baggrunde og ekspertiseniveauer få adgang til og fortolke oplysninger om deres funktion og resultater. Denne tilgængelighed sikrer, at AI-teknologier ikke kun er forståelige, men også brugbare for en bred vifte af brugere, herunder dem med handicap eller begrænset teknisk viden. I udviklingen af AI-drevne tilgængelighedsværktøjer til mennesker med handicap gør gennemsigtighed det f.eks. muligt for brugerne at forstå, hvordan værktøjerne fungerer, og hvordan de kan få gavn af dem.

Et illustrativt eksempel på vigtigheden af gennemsigtighed i AI-systemer er udviklingen af AI-modeller til medicinsk diagnose, f.eks. til at opdage kræft. Selv om en AI-model er meget nøjagtig med en succesrate på 99 %, kan den resterende fejlmargen på 1 % have livstruende konsekvenser for patienterne. I sådanne kritiske scenarier bliver gennemsigtighed afgørende for at sikre, at sundhedspersonalet kan forstå, hvordan AI-modellen nåede frem til sin diagnose, og kan verificere dens nøjagtighed, før de træffer beslutninger om behandling. Ved at give gennemsigtige forklaringer på sin beslutningsproces bliver AI-modellen et værdifuldt værktøj for sundhedspersonalet, der forbedrer deres evne til at diagnosticere og behandle patienter effektivt.

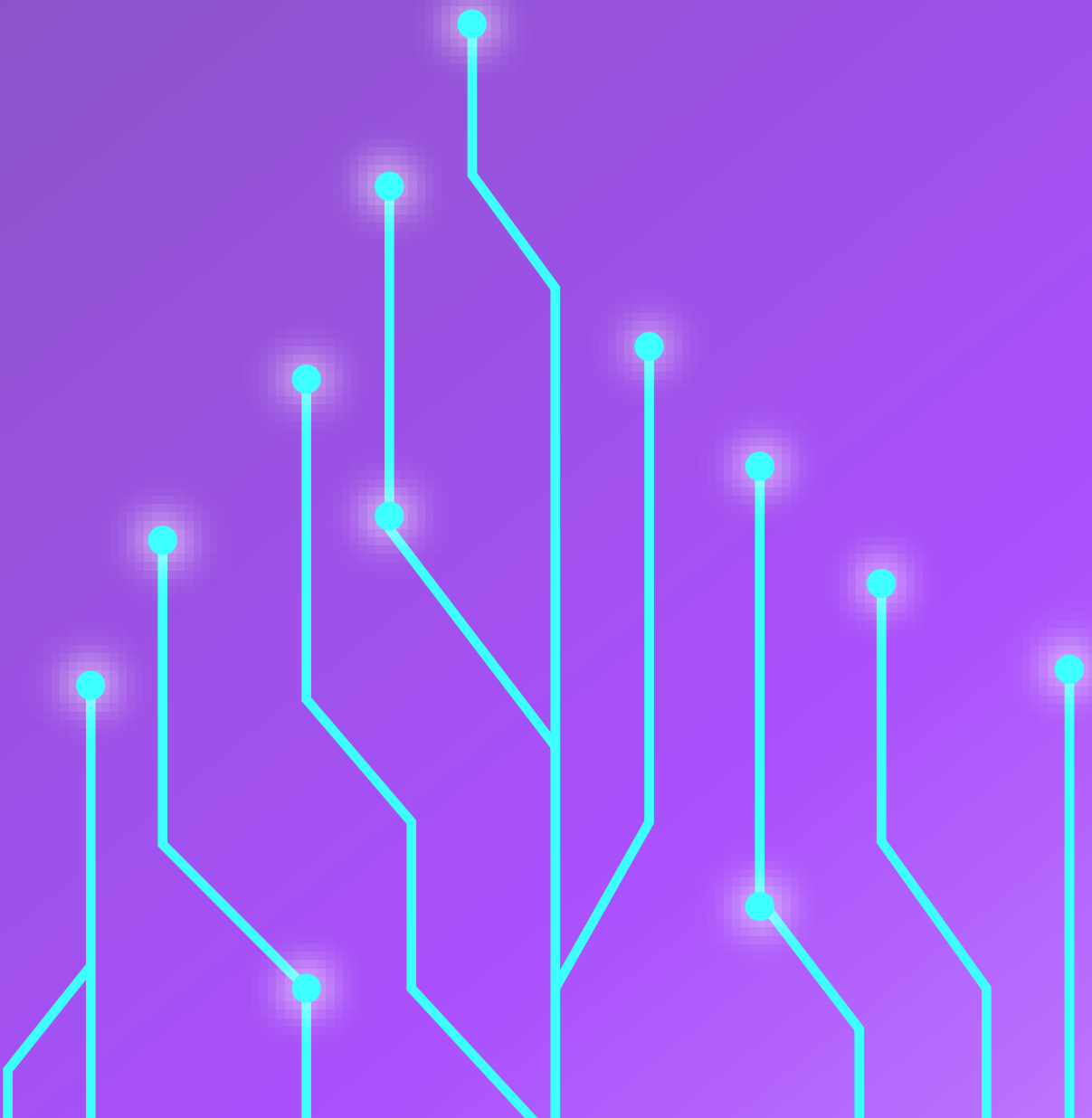
Som tidligere nævnt i dette hæfte, refererer algoritmisk bias til systematiske fejl eller uretfærdighed i AI-algoritmer, der resulterer i diskriminerende resultater for visse personer eller grupper. Disse bias kan opstå fra forskellige kilder, herunder forudindtagede træningsdata, mangelfuldt algoritmisk design eller menneskelige bias, der er kodet ind i systemet. Konsekvenserne af algoritmisk bias kan være vidtrækkende, fastholde uligheder, forstærke stereotyper og underminere tilliden til AI-systemer.

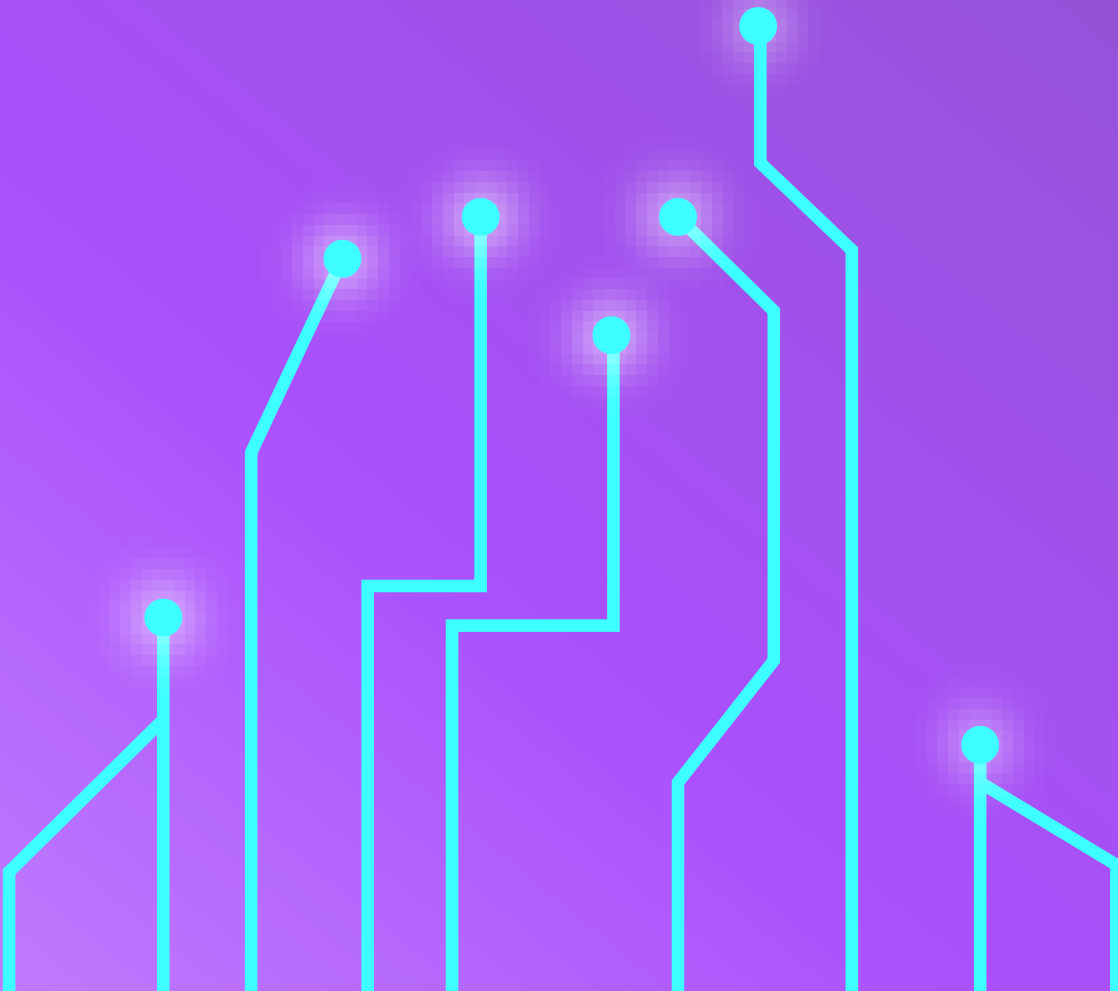




03. Forholdet mellem gennemsigtighed og algoritmisk bias

CU4 | Gennemsigtighed





03. Forholdet mellem gennemsigtighed og algoritmisk bias

Uigennemsigtighed eller mangel på gennemsigtighed forværrer de risici, der er forbundet med algoritmisk bias.

Ofte er AI-algoritmer uigennemsigtige, hvilket betyder, at forklaringer på deres beslutninger og handlinger ikke er umiddelbart tilgængelige for alle interessenter. Denne uigennemsigtighed kan stamme fra forskellige kilder, herunder institutionel hemmeligholdelse, virksomhedsfortrolighed eller teknisk kompleksitet. Når interessenter ikke har adgang til information om AI-systemer, er de ikke i stand til at vurdere retfærdigheden, pålideligheden eller de etiske konsekvenser af algoritmiske resultater, hvilket fører til manglende ansvarlighed og potentiel skade.

Gennemsigtighed er en vigtig modgift til uigennemsigtighed i AI-systemer, så interessenter kan granske og udfordre algoritmiske beslutninger og dermed mindske risikoen for algoritmisk bias. Ved at øge gennemsigtigheden kan AI-udviklere og -udøvere give interessenter indsigt i, hvordan AI-systemer fungerer, hvorfor visse beslutninger træffes, og hvilke faktorer der påvirker deres output. Transparente AI-systemer giver interessenter mulighed for at identificere og håndtere bias, validere algoritmisk nøjagtighed og holde udviklere ansvarlige for den etiske og retfærdige brug af AI-teknologier.

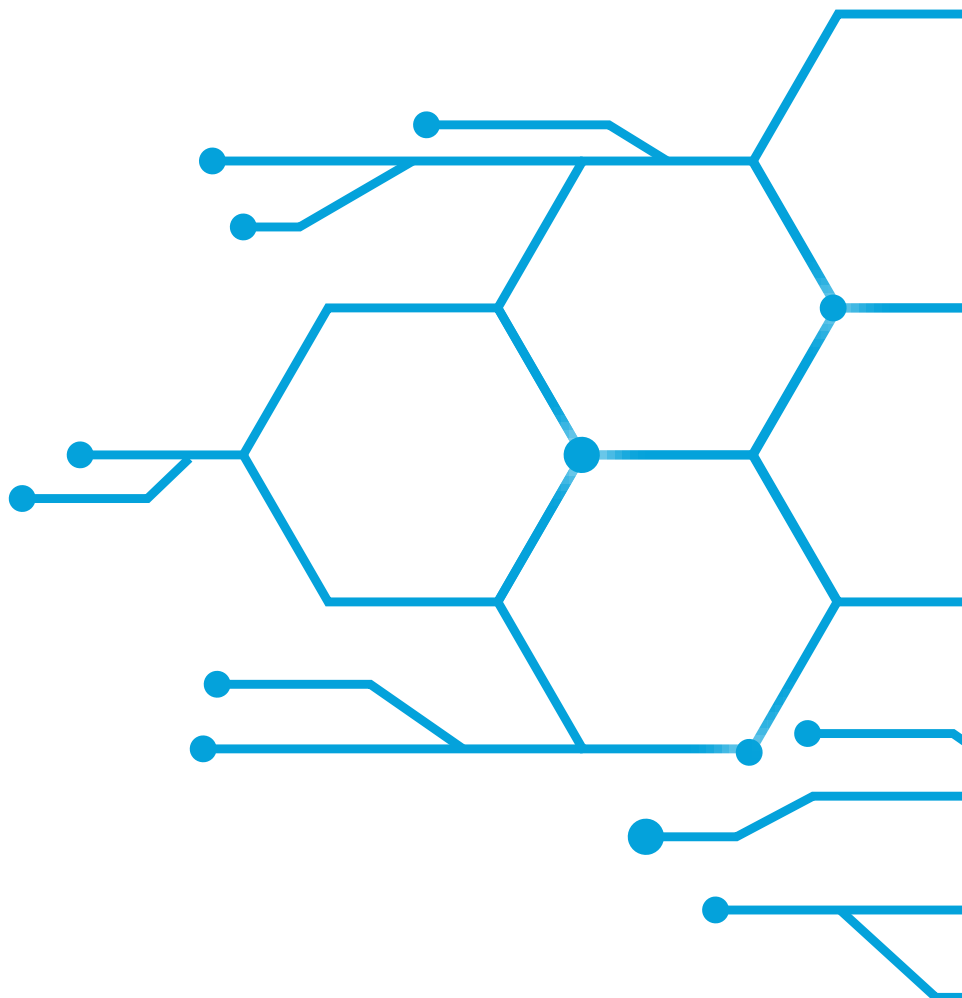


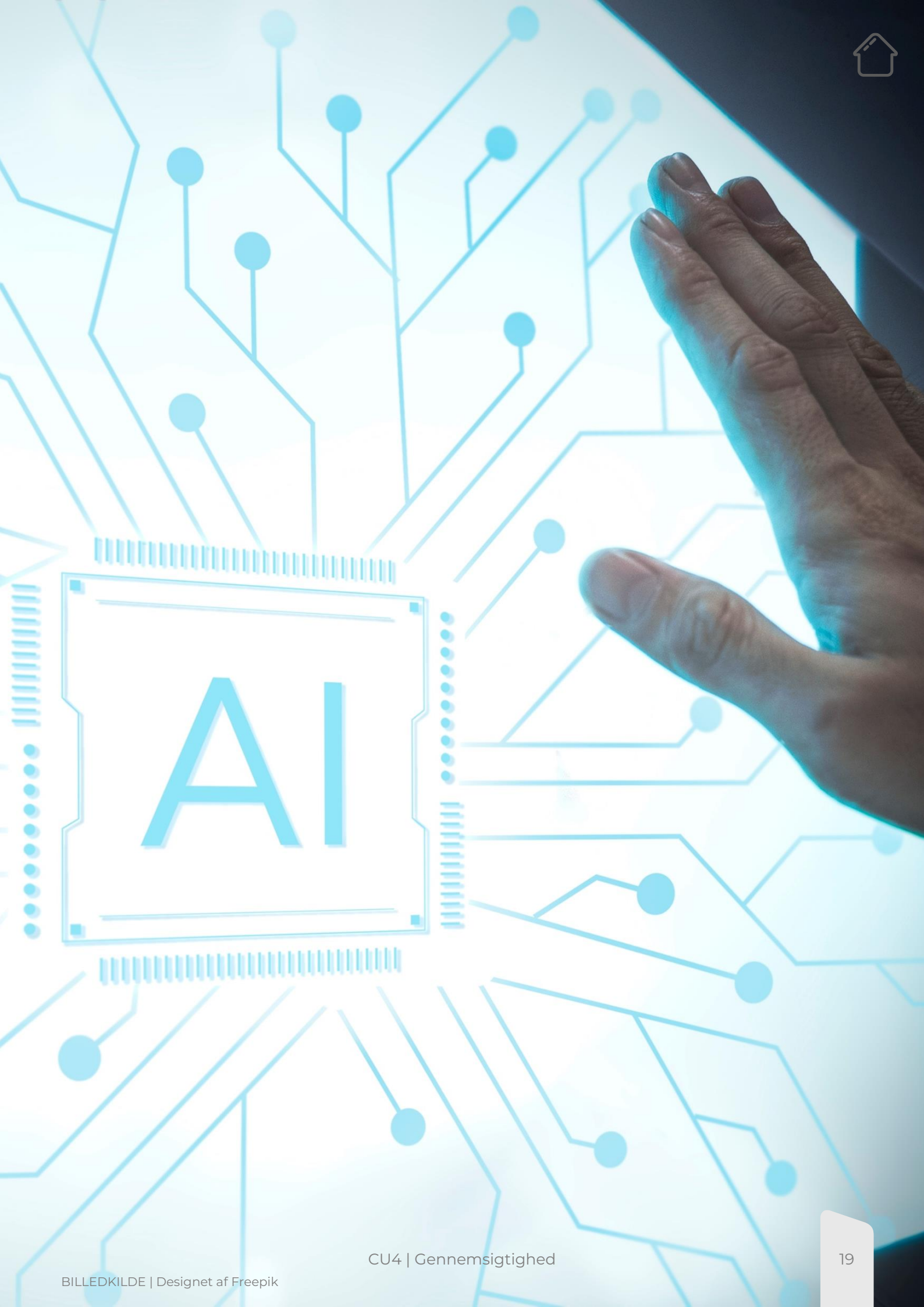
En af de vigtigste fordele ved gennemsigtighed i forbindelse med algoritmisk bias er muligheden for at opdage og afbøde forudindtagede resultater. Når AI-algoritmer er gennemsigtige, kan interessenter undersøge beslutningsprocessen og identificere tilfælde, hvor der kan være bias til stede. For eksempel i forbindelse med et AI-system til ansættelse gør gennemsigtighed det muligt for interessenter at vurdere, om systemet uretfærdigt diskriminerer visse demografiske grupper i udvælgelsesprocessen. Ved at identificere forudindtagede resultater kan interessenter træffe korrigerende foranstaltninger for at mindske skaden forårsaget af algoritmisk bias og fremme fairness og retfærdighed.

Desuden fremmer gennemsigtighed ansvarlighed og tillid til AI-systemer. Når interessenter har adgang til information om AI-algoritmer, kan de holde udviklere og udøvere ansvarlige for den etiske og retfærdige brug af AI-teknologier. Gennemsigtige AI-systemer opbygger tillid blandt brugere, lovgivere og den brede offentlighed og fremmer tilliden til pålideligheden og retfærdigheden af algoritmiske resultater. For eksempel i forbindelse med implementering af AI-systemer i strafferetten eller sundhedsvæsenet gør gennemsigtighed det muligt for interessenter at forstå, hvordan beslutninger træffes, og at sikre, at disse beslutninger er i overensstemmelse med etiske principper og juridiske standarder.



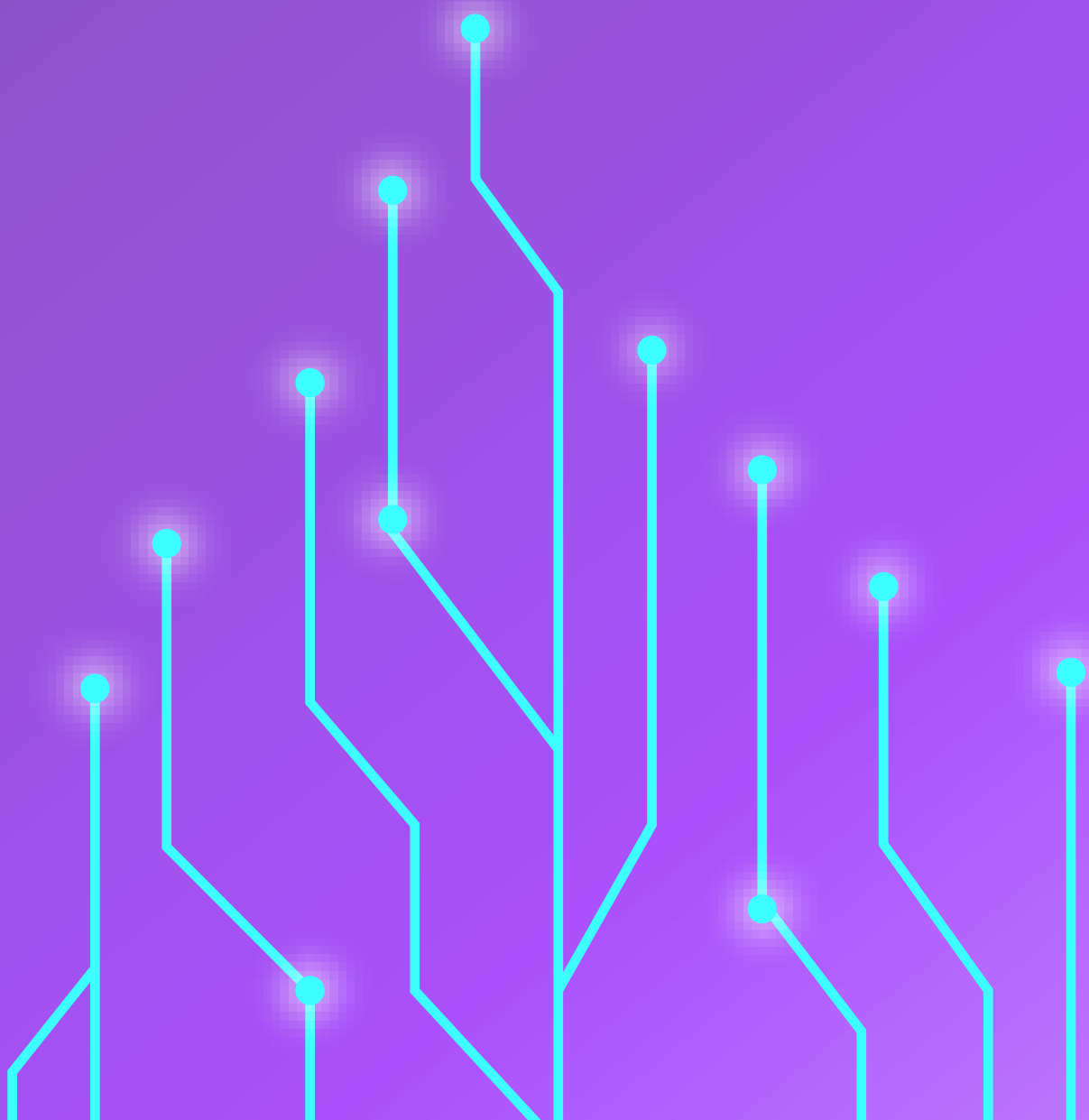
Gennemsigtighed spiller en afgørende rolle for at håndtere og mindske algoritmisk bias i AI-systemer. Ved at øge gennemsigtigheden kan interessenter opdage og afbøde forudindtagede resultater, fremme ansvarlighed og opbygge tillid til AI-teknologier. Efterhånden som AI fortsætter med at udvikle sig og blive mere integreret i forskellige aspekter af samfundet, er gennemsigtighed fortsat afgørende for at sikre, at AI-systemer udvikles og implementeres på en måde, der opretholder etiske standarder og fremmer retfærdighed og lighed. Gennem en omfattende forståelse af forholdet mellem gennemsigtighed og algoritmisk bias kan de studerende bidrage til en ansvarlig og etisk udvikling af AI-teknologier, og dermed skabe en mere retfærdig og inkluderende fremtid.

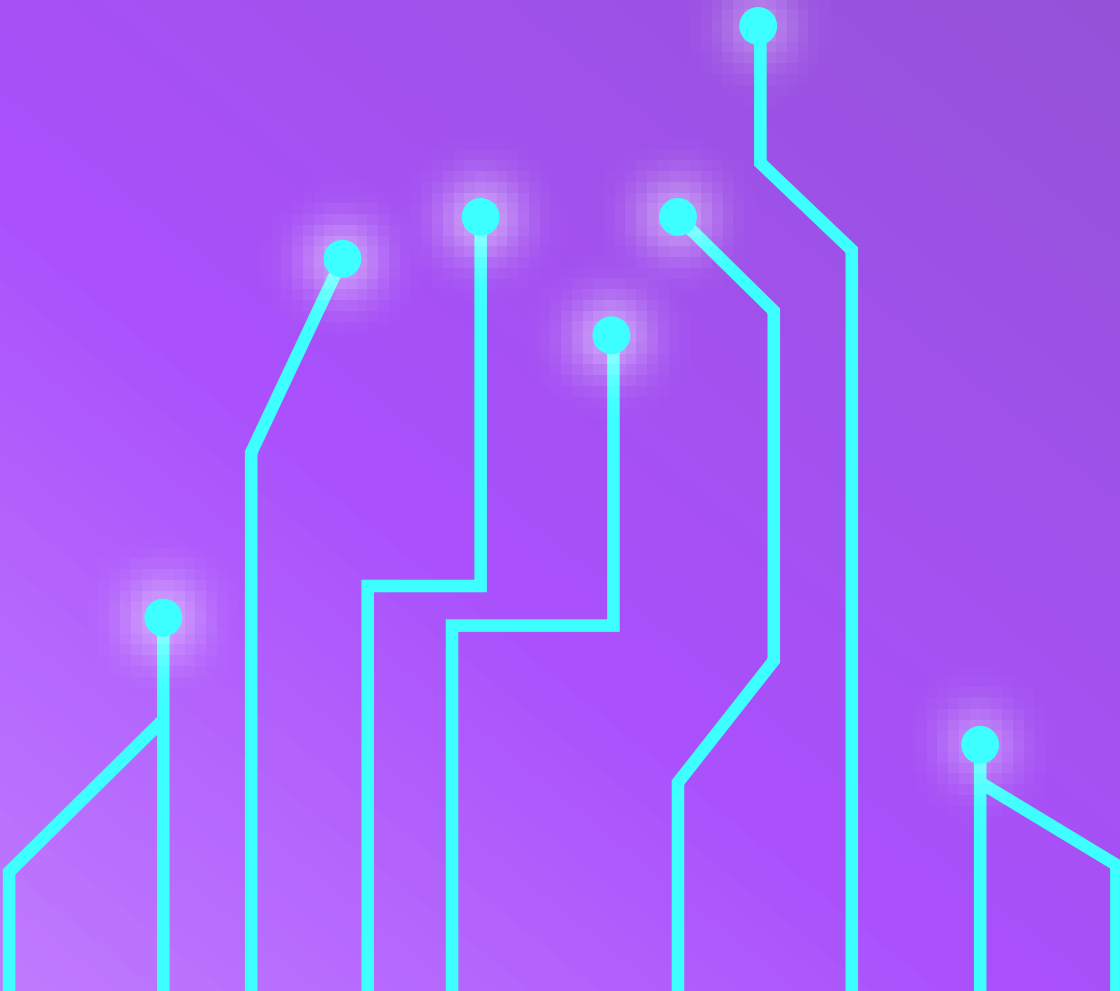




04. Strategier til fremme af gennemsigtighed i AI-systemer

CU4 | Gennemsigtighed





04. Strategier til fremme af gennemsigtighed i AI-systemer

Der er forskellige strategier til at fremme gennemsigtighed i systemer med kunstig intelligens (AI), f.eks. ved at bruge fortolkelige modeller, give klar dokumentation og kommunikere beslutningsprocesserne.

> Fortolkelige modeller

Fortolkelige modeller er en vigtig strategi for at fremme gennemsigtighed i AI-systemer. Det er maskinlæringsmodeller, som producerer resultater, der er lette at forstå og fortolke for mennesker. Her er nogle eksempler:

- **Lineær regression:** Lineær regression er en enkel og fortolkelig model, der ofte bruges til at forudsige numeriske resultater. Den fungerer ved at tilpasse en lige linje til datapunkterne, hvilket gør det nemt at fortolke forholdet mellem inputvariablerne og outputtet.
- **Beslutningstræer:** Beslutningstræer er hierarkiske modeller, der træffer beslutninger baseret på en række hvis-så-sætninger. Hver node i træet repræsenterer en beslutning baseret på en funktion i dataene, hvilket gør det nemt at følge logikken bag modellens forudsigelser.



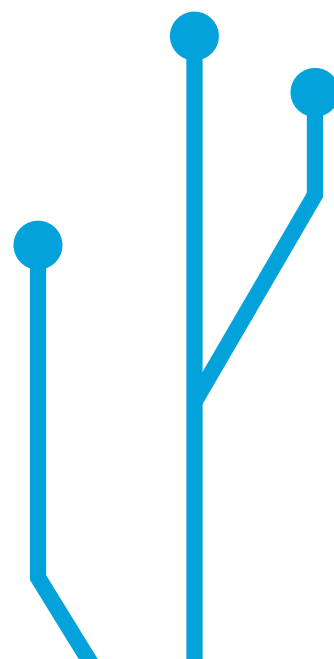
- **Logistisk regression:** Logistisk regression er en statistisk model, der bruges til binære klassificeringsopgaver. Den beregner sandsynligheden for, at en forekomst tilhører en bestemt klasse baseret på dens inputfunktioner, hvilket gør den fortolkelig og let at forstå.
- **Regelbaserede modeller:** Regelbaserede modeller, som f.eks. klassifikations- og regressionstræer (CART) eller beslutningsregler, oversætter direkte inputfunktioner til beslutningsregler. Disse regler er nemme at fortolke og kan give indsigt i, hvordan modellen laver forudsigelser.
- **Generaliserede additive modeller (GAM):** GAM'er er fleksible modeller, der kan indfange komplekse forhold mellem inputvariabler og målvariablen og samtidig bevare fortolkningen. De bruger glatte funktioner til at repræsentere forholdet mellem hver inputvariabel og output, hvilket giver mulighed for nem fortolkning af modellens forudsigelser.

I modsætning til komplekse black box-modeller giver fortolkelige modeller interessenter mulighed for at forstå, hvordan AI-algoritmer træffer beslutninger, og hvilke faktorer der påvirker deres output. Ved at bruge fortolkelige modeller kan AI-udviklere forbedre gennemsigtigheden og ansvarligheden, så interessenter kan validere algoritmiske resultater og identificere potentielle bias eller fejl. I forbindelse med et AI-kreditscoringssystem giver brug af fortolkelige modeller for eksempel interessenter mulighed for at forstå de faktorer, der bidrager til kreditbeslutninger, såsom indkomst, kredithistorik og gældsniveauer, hvilket fremmer gennemsigtighed og retfærdighed i udlånspraksis.

➤ Tydelig dokumentation

Tydelig dokumentation er en anden vigtig strategi for at fremme gennemsigtighed i AI-systemer. Dokumentation giver interessenter indsigt i design, udvikling og implementering af AI-algoritmer, herunder datakilder, forbehandlingsteknikker, modelarkitekturer og evalueringsmålinger.

Ved at dokumentere AI-systemer grundigt kan udviklere øge gennemsigtigheden og ansvarligheden, så interessenter kan forstå de underliggende processer og antagelser i AI-teknologier. For eksempel i udviklingen af et AI-system til forudsigelig vedligeholdelse af industrielt udstyr giver klar dokumentation interessenter mulighed for at vurdere pålideligheden og nøjagtigheden af forudsigelige modeller, forstå vedligeholdelses anbefalinger og verificere overholdelse af sikkerhedsstandarder.





> **Effektiv kommunikation af beslutningstagning**

Effektiv kommunikation af beslutningsprocesser er afgørende for at fremme gennemsigtighed i AI-systemer. Kommunikation sikrer, at interessenter er informeret om rationalet, logikken og konsekvenserne af AI-algoritmiske beslutninger.

Ved at kommunikere beslutningsprocesserne klart og tydeligt kan AI-udviklere opbygge tillid blandt brugere, myndigheder og den brede offentlighed. For eksempel i forbindelse med implementering af AI-systemer til sundhedsdiagnoser sikrer effektiv kommunikation, at sundhedspersonale og patienter forstår, hvordan diagnostiske beslutninger træffes, så de kan stole på og verificere nøjagtigheden af AI-genererede diagnoser.



> **Inddragelse og bemyndigelse af AI-interessenter**

Inddragelse og bemyndigelse af AI-interessenter i hele AI-livscyklussen er afgørende for at sikre gennemsigtighed og ansvarlighed. Inddragelse af interessenter involverer brugere, kunder, medarbejdere, ledere, lovgivere og samfundet i design, udvikling, implementering og evaluering af AI-systemer.

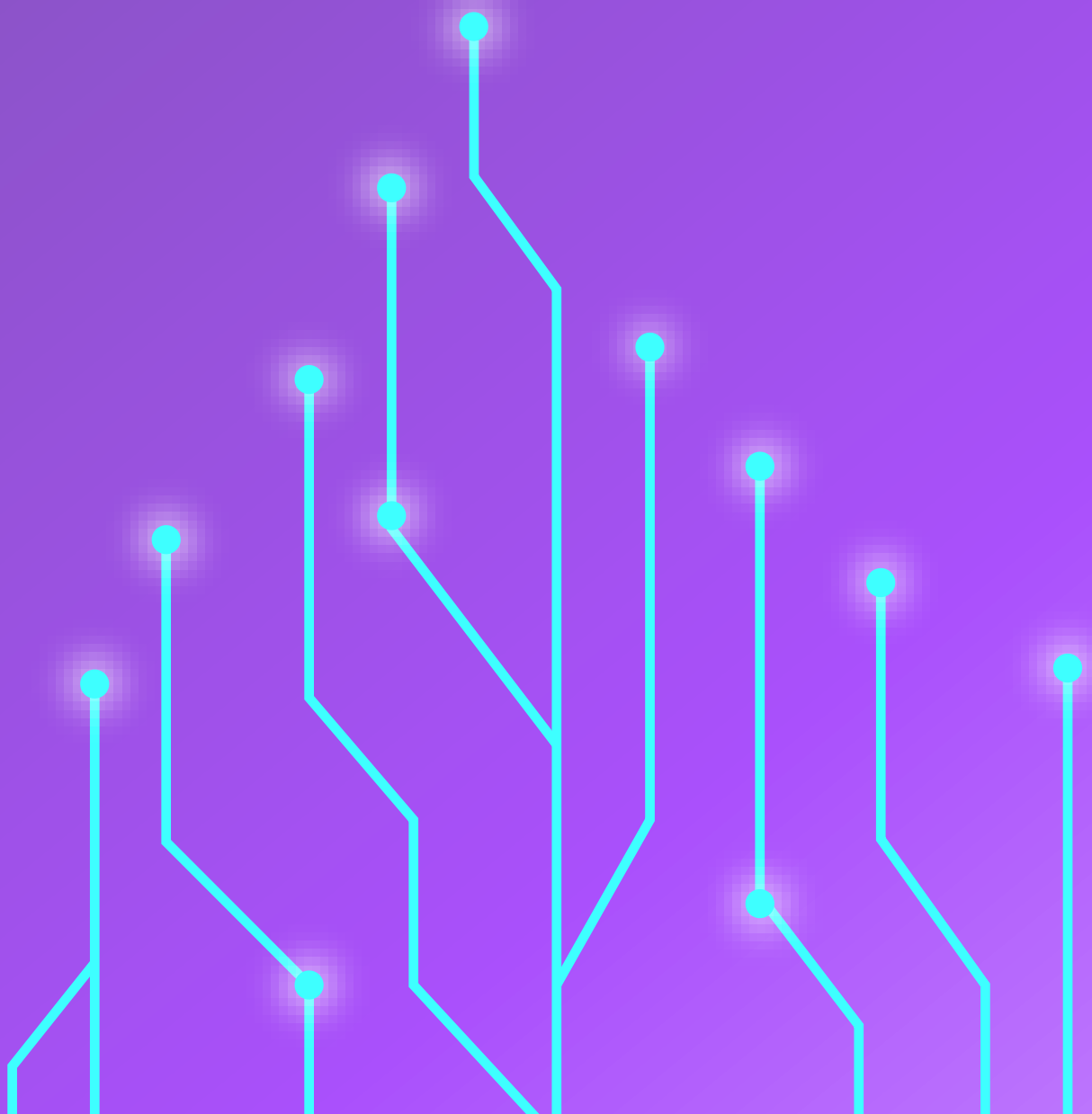
Ved at inddrage interessenter kan AI-udviklere få værdifuld indsigt i deres behov, præferencer og bekymringer og dermed fremme gennemsigtighed, ansvarlighed og etisk beslutningstagning. I udviklingen af AI-drevne autonome køretøjer sikrer inddragelse af myndigheder og samfund for eksempel, at der tages højde for sikkerhed, privatliv og etiske overvejelser, hvilket øger gennemsigtigheden og tilliden til teknologien.

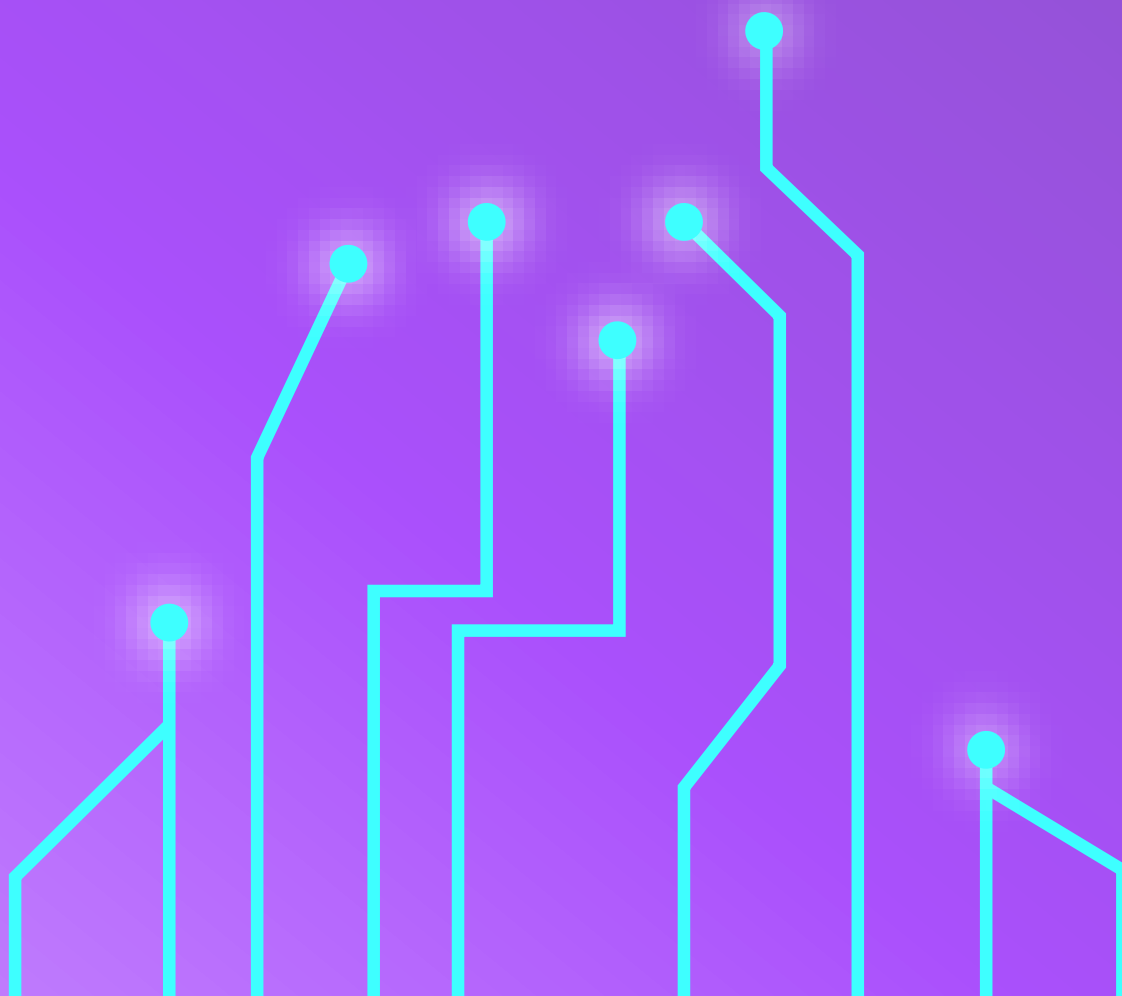




05. Konklusion

CU4 | Gennemsigthed

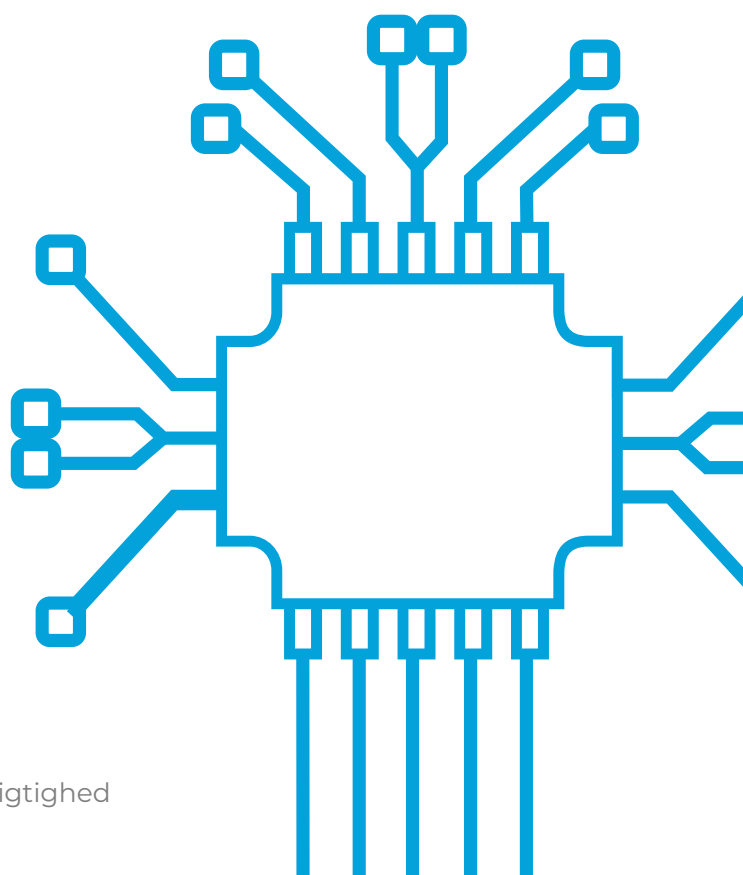


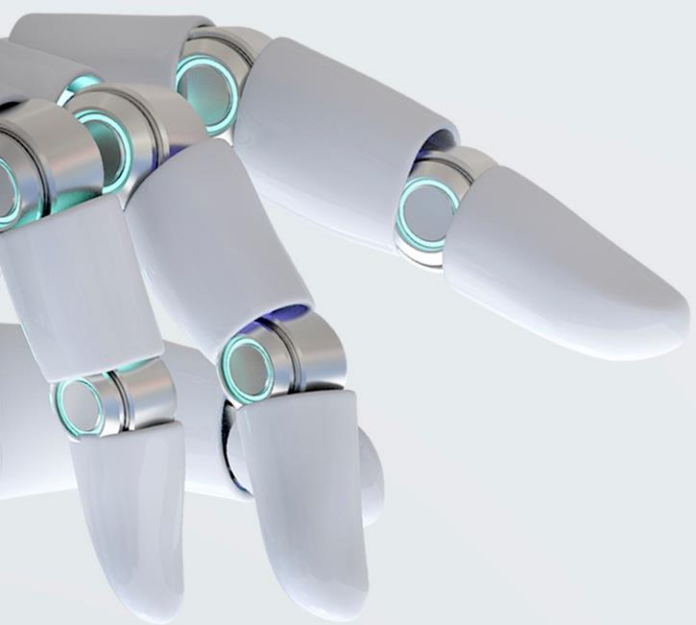


05. Konklusion

Konklusionen er, at vigtigheden af gennemsigtighed i AI-systemer ikke kan overvurderes, da det danner grundlaget for at opbygge tillid, ansvarlighed og mindske algoritmisk bias. Desuden fremhæver forståelsen af forholdet mellem gennemsigtighed og algoritmisk bias, behovet for at adressere uigennemsigtighed som et middel til at identificere, forhindre og afbøde forudindtagede resultater i AI-systemer.

Endelig giver udforskningen af strategier til fremme af gennemsigtighed de studerende praktiske værktøjer til at øge ansvarligheden og fremme tilliden til AI-teknologier. Gennem en omfattende forståelse af disse begreber er de studerende bedre forberedt på at navigere i de etiske udfordringer ved udvikling og anvendelse af kunstig intelligens, hvilket bidrager til udviklingen af ansvarlige og retfærdige AI-systemer.







Charlæ



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