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

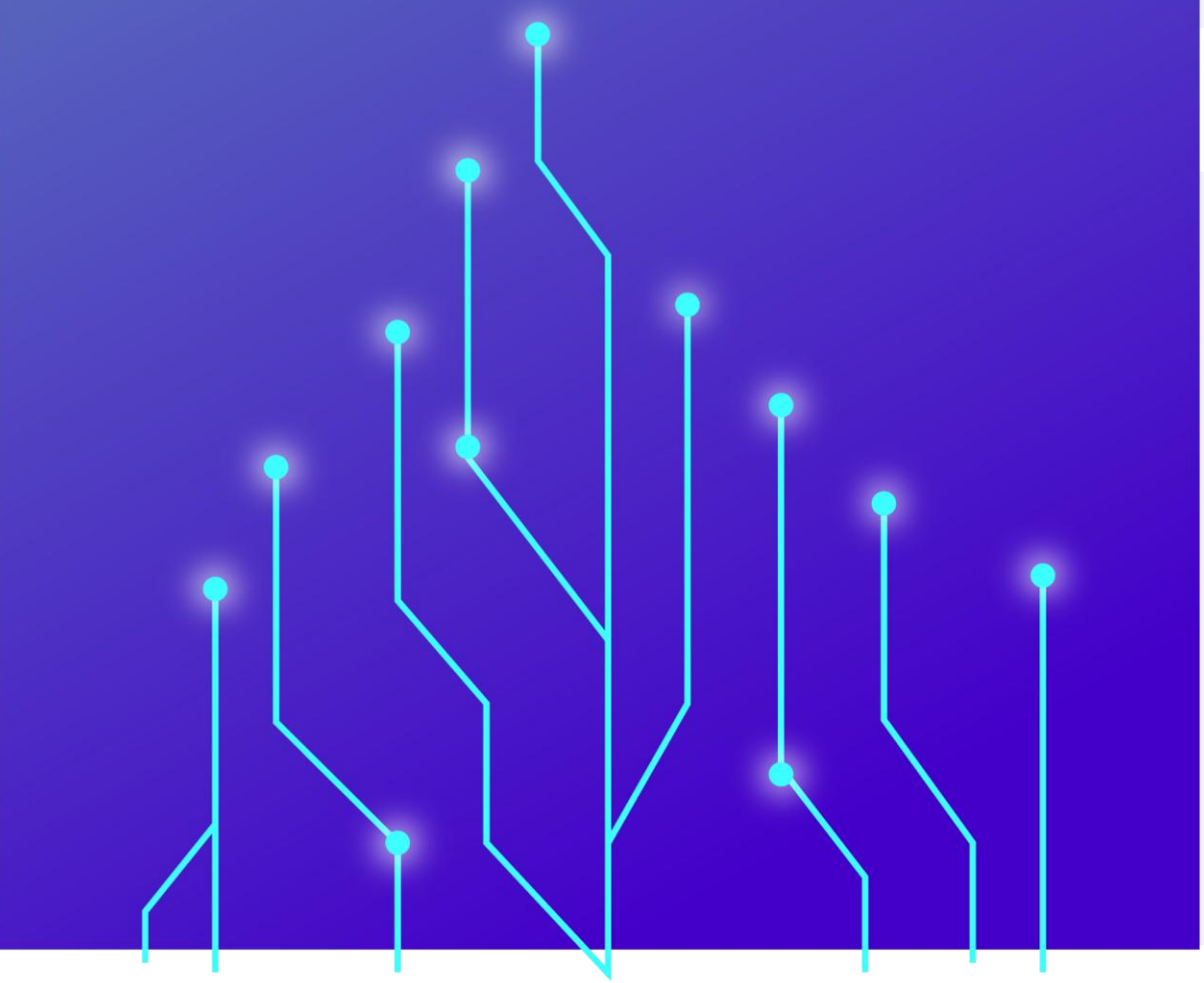
## Håndbog

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

# Indholdsfortegnelse

Comentado [KM1]: Must be updated after proof reading is done.

# CU1 | AI-etik - en praktisk tilgang



# Indhold

Comentado [KM2]: Must be updated in the very end.

# 1. Indledning

## TIPS OG ANBEFALINGER TIL UNDERVISER

Denne håndbog har til formål at guide dig gennem materialet til kurset. De understøttende PowerPoint-filer indeholder den samme information som den, der blev gennemgået i e-læringsdelen, men i mere detaljeret form. PowerPoint-filerne er omfattende og er ikke beregnet til at blive gennemgået i deres helhed under de interaktive sessioner. I stedet kan du spørge i begyndelsen af de interaktive sessioner, hvilke emner der var sværest at forstå for de studerende, og kun bruge disse supportslides i forløbet. Du kan spørge om dette ved eksempelvis at bruge værktøjer som Mentimeter eller Kahoot og tilføje kursusemnerne fra kompetenceenhedens indhold, der er beskrevet i punktform i slutningen af introduktionsafsnittet. Desuden er det meningen, at casestudier, som er relateret til de enkelte kursusdele, også skal gennemgås i den interaktive session. Alle yderligere forslag i håndbogen er kun vejledende; du er velkommen til at bruge dem, som du finder det passende.



Hvad ville der ske, hvis kunstig intelligens (AI) fik magt til at beslutte, om du får et lån? Eller hvad sker der, hvis AI skal dømme, hvem der skal i fængsel? Hvad er konsekvenserne, hvis en selvkørende bil skal beslutte, om den skal redde en fodgænger eller livet for dem, der sidder i bilen? Hvad hvis et AI-baseret rekrutteringssystem skulle beslutte, om en kvindelig eller mandlig kandidat skulle ansættes som flykaptajn?

Den stigende integration af kunstig intelligens i forskellige aspekter af samfundet giver anledning til betydelige etiske bekymringer. Fra beslutninger om lån, strafudmåling og selvkørende køretøjer til ansættelsespraksis rejser afhængigheden af AI-algoritmer spørgsmål om etiske principper som retfærdighed, gennemsigtighed og ansvarlighed. AI-teknologier har potentiale til at fastholde fordomme, diskrimination og uligheder, hvis de ikke udvikles og reguleres omhyggeligt. At sikre, at AI-systemer prioriterer etiske overvejelser, er afgørende for at fremme retfærdighed, lighed og tillid i systemernes anvendelse på tværs af forskellige domæner. (UNESCO, 2023c)

Denne enhed vil gå dybere ind i etiske principper, rammer og retningslinjer, der påvirker udviklingen af AI-løsninger. AI-udviklingens livscyklus undersøges fra problemformulering til drift og overvågning af AI-løsningen. Relaterede principper og inddragelse af interessenter overvejes i hvert trin af processen. På indgående vis behandles i denne kursusenhed følgende emner.

- AI-etik og dens betydning for udvikling af AI-løsninger
- Etiske principper for AI

- Etiske rammer, retningslinjer og værktøjer
- Ekspertgruppe på højt niveau; troværdige AI-rammer
- EU's lov om kunstig intelligens
- Livscyklus for AI-udvikling
- Inddragelse af interessenter i AI-udviklingens livscyklus
- AI-udviklingens livscyklus med tilhørende etiske principper og interessenter

## 2. Definition af etiske principper, rammer og retningslinjer

Kunstig intelligens (AI) påvirker alle aspekter af livet, både arbejde og fritid, og tilbyder løsninger på globale problemer som klimaforandringer, og AI kan også have indflydelse på adgangen til sundhedsydelser. I lyset af AI's udbredte integration i økonomi og samfund opstår spørgsmålet om passende politiske og institutionelle rammer, der er nødvendige for at styre AI's udvikling og anvendelse og sikre samfundsmæssig velfærd.

Brede etiske overvejelser inden for AI har ført til skabelsen og offentliggørelsen af adskillige tilgange til at sikre en etisk anvendelse af kunstig intelligens (Prem, 2023). Etiske principper, rammer og retningslinjer er almindeligt anvendt og bredt anerkendt som kernestrukturen i etisk beslutningstagning på tværs af mange discipliner. De giver en omfattende tilgang til at navigere i etiske spørgsmål, fra etablering af grundlæggende værdier til vejledning i specifikke handlinger. Selv om de ofte er sammenflettede, kan de beskrives på følgende måde:



**Etiske principper** giver en moralsk ramme for vurdering og løsning af etiske dilemmaer. Etiske principper udgør de kerneværdier og overbevisninger, der fungerer som et kompas for etisk beslutningstagning og giver et grundlag for at afgøre, hvad der er rigtigt, og hvad der er forkert i forskellige sammenhænge (Prem, 2023; Zhou & Chen, 2022).

**Etiske rammer** er systematiske tilgange eller modeller, der hjælper enkeltpersoner og organisationer med at navigere i etiske spørgsmål. Etiske rammer er omfattende modeller eller strukturerede metoder, der gør det lettere at undersøge og løse etiske problemer. De giver en ramme for at vurdere de etiske konsekvenser af en handling og guider beslutningsprocessen, så det sikres, at der systematisk tages hensyn til etiske overvejelser (Prem, 2023; Zhou & Chen, 2022).

**Etiske** retningslinjer er specifikke regler, standarder og *best practice*, der giver praktisk vejledning i etisk adfærd i forskellige sammenhænge. Etiske retningslinjer er detaljerede regler og benchmarks, der giver praktiske råd om, hvordan man anvender etiske principper i forskellige scenarier. Disse

retningslinjer hjælper både enkeltpersoner og organisationer til effektivt at håndtere etiske problemer og til at fastholde en etisk adfærd i deres daglige drift og beslutningstagning (Prem, 2023; Zhou & Chen, 2022).

## 3. Fælles indsats for at forme det etiske landskab

Etiske principper, rammer og retningslinjer kommer fra mange forskellige områder og organisationer, herunder private virksomheder, forskningsinstitutioner og organisationer i den offentlige sektor, og udgør en fælles indsats for at forme det etiske landskab inden for kunstig intelligens (Jobin et al., 2019; Morley et al., 2020).

Vejledningsdokumenter og rapporter om etisk AI er eksempler på ikke-lovgivningsmæssige politiske værktøjer eller "blød lovgivning". I modsætning til "hård lovgivning", som omfatter lovbestemmelser, der er fastsat af lovgivende organer for at påbyde eller forbyde visse former for adfærd, har etiske retningslinjer ingen juridisk gyldighed, men de har overtalelseskraft. Disse dokumenter er designet til at understøtte beslutningsprocesser inden for specifikke områder og har vist sig at have stor indflydelse på praksis (Jobin et al., 2019).

**Eksempler på dokumenter, der definerer de etiske principper for kunstig intelligens, er:**

OECD's AI-principper fremmer innovativ og troværdig brug af AI, der respekterer menneskerettigheder og demokratiske værdier. De blev vedtaget i maj 2019 og sætter standarder for kunstig intelligens, der sigter mod at være praktiske og fleksible. (OECD, 2019)

Beijing Academy of Artificial Intelligence (BAAI) udgav "Beijing AI Principles", som er en oversigt, der skal guide forskning og udvikling, implementering og styring af kunstig intelligens ("Beijing Academy of Artificial Intelligence - Beijing AI Principles", 2019; *Beijing Artificial Intelligence Principles*, 2022).

Store brancheaktører som Google, IBM, Microsoft og Intel har udviklet deres egne etiske retningslinjer, der afspejler deres unikke perspektiver og arbejdsområder (Jobin et al., 2019; Morley et al., 2020).

Regeringsorganer er ikke blevet ladt i stikken, men har fået vigtige dokumenter som f.eks. Montreal-erklæringen og bidrag fra organer som House of Lords Select Committee on Artificial Intelligence og Europa-Kommissionens Ekspertgruppe, der alle er involveret i AI-etisk regulering og politik (Morley et al., 2020).

Akademiske institutioner og tænketanke, herunder Future of Life Institute, IEEE og AI4People, har uddybet debatten med videnskabelig forskning og teoretiske modeller (Floridi et al., 2018; Jobin et al., 2019; Morley et al., 2020).

Kombinationen af disse bestræbelser afspejler en global bevægelse mod ansvarlig AI, og hvert bidrag udgør en brik i det større puslespil om at opbygge etiske AI-systemer, der er i overensstemmelse med samfundets værdier og normer.

## TIPS OG ANBEFALINGER TIL UNDERVISERE

Du kan bede eleverne om at tjekke de ovennævnte kilder eller nogle af dem og diskutere forskellene i de principper, de præsenterer.

## 4. Etiske principper

Landskabet af etiske principper inden for kunstig intelligens er rigt og mangfoldigt med mange principper foreslået af en række forskellige aktører, herunder akademiske institutioner, brancheledere, regeringsorganer og internationale organisationer. Udbredelsen af disse principper afspejler en voksende konsensus om behovet for etisk vejledning i udviklingen og anvendelsen af teknologier inden for kunstig intelligens (Jobin et al., 2019; Morley et al., 2020; Prem, 2023).

Jobin et al. (2019) lavede en omfattende gennemgang af forskellige etiske AI-principper og destillerede dem for at identificere fælles temaer og elementer, der går igen i forskellige kilder. Blandt de retningslinjer, de gennemgik, fandt de en begyndende konsensus om vigtige etiske principper for kunstig intelligens. Selvom de ikke støttes universelt, omfatter de konvergerende principper åbenhed, fairness og retfærdighed, ansvar, erstatningsansvar, privatlivets fred, gavnlighed, frihed og autonomi, tillid, bæredygtighed, menneskelig værdighed og solidaritet.



De identificerede, fælles elementer henviser til en række fælles betænkeligheder og bestræbelser i det globale samfund. De peger i retning af en begyndende konsensus om de værdier, som mange mener bør understøtte udviklingen og anvendelsen af systemer med kunstig intelligens. Denne foreløbige enighed udgør et fælles grundlag, der kan fungere som udgangspunkt for at opstille forventninger og evaluere resultater (Morley et al., 2020).

Morley et al. (2020) anerkender denne kollektive enighed som et grundlæggende udgangspunkt. De hævder, at disse principper, hvorom der er enighed, karakteriserer en etisk tilpasset AI på følgende måde:

- Gavnlig og respektfuld over for mennesker og miljø (gavnlighed)
- Robust og sikker (ikke-skadelighed)
- Respekt for menneskelige værdier (autonomi)
- Fair (retfærdighed)
- Forklarlig, ansvarlig og forståelig (forklarlighed)

## Gavnlighed

Princippet om gavnlighed er en af de vigtigste etiske overvejelser i udviklingen af kunstig intelligens, og det understreger vigtigheden af, at systemer med kunstig intelligens ikke kun er til gavn for enkeltpersoner og samfundet, men også for miljøet. Ifølge Floridi et al. (2018) er målet at sikre, at kunstig intelligens fungerer på en måde, der fremmer generel trivsel og miljømæssig sundhed.

Jobin et al. (2019) opregner flere vigtige tiltag for effektivt at fremme kunstig intelligens:

- Tilpas AI til menneskelige værdier, og sørg for, at AI-systemer understøtter og forstærker disse værdier i stedet for at underminere dem.
- Inddrag AI-interessenter i udviklingsprocessen og tag hensyn til deres behov og feedback for at forbedre AI-systemernes indvirkning på menneskers liv.
- Arbejd aktivt på at minimere og løse potentielle interessekonflikter ved at bruge brugerfeedback til at styre udviklingen og anvendelsen af AI.
- Skab nye måder at måle menneskelig trivsel på.



## TIPS OG ANBEFALINGER TIL UNDERVISERE

Diskuter de miljømæssige og samfundsmæssige konsekvenser med eleverne. For eksempel kan TikTok-algoritmer vise skadeligt indhold, hvis brugeren viser interesse for det, hvilket kan give negative effekter på den mentale sundhed. Derudover kan den betydelige mængde elektricitet, der bruges til at skabe store sprogmodeller, efterlade et betydeligt CO2-aftryk.

## Ikke-skadelighed

Det etiske princip om ikke-skadelighed (*non-maleficence*) handler om at forebygge skade. Ifølge Floridi et al. (2018) er ikke-skadelighed ensbetydende med proaktive foranstaltninger mod både forsætligt misbrug fra menneskers side og utilsigtede negative handlinger fra AI-systemer, der kan påvirke menneskelig adfærd. Det centrale mål er at sikre, at AI ikke fører til skadelige virkninger uanset deres oprindelse.

For at håndtere risici og beskytte mod skade i AI-systemer har Jobin et al. (2019) identificeret følgende tilgange:

- Brug af strategisk risikostyring til at forudse og reducere potentielle risici i forbindelse med AI-systemer.

- Implementering af en kombination af tekniske foranstaltninger og styringspolitikker, der dækker hele livscyklussen for AI, med det formål at forhindre skade, før den opstår.
- Visse skader kan være uundgåelige på trods af alle anstrengelser, hvilket kræver en omfattende risikovurdering. Dette omfatter foranstaltninger til at reducere og afbøde risici og klart placere et ansvar, når der opstår skade.

## TIPS OG ANBEFALINGER TIL UNDERVISERE

Diskuter med de studerende de mulige skader forårsaget af AI. F.eks. viste Amazons AI-rekrutteringsværktøj bias mod kvindelige ansøgere. Se efter flere detaljer, f.eks. her: Insight - Amazon skrotter hemmeligt AI-rekrutteringsværktøj, der viste bias mod kvinder | Reuters

### Autonomi

Autonomi i forbindelse med AI fokuserer på at opretholde en balance mellem menneskelig beslutningstagning og AI's indflydelse. Mennesker skal bevare beslutningskontrollen og vælge, hvornår de vil træffe deres egne valg, og hvornår de vil lade AI'en bestemme. Men mennesker skal altid have mulighed for at tage kontrollen tilbage fra AI'en, hvis det er nødvendigt, hvilket betyder, at de har det sidste ord (Floridi et al., 2018).



Jobin et al. (2019) identificerer en række tiltag til fremme af frihed og autonomi i AI-systemer:

- Forbedring af gennemsigtigheden og forudsigeligheden af AI-systemer, så brugerne kan forstå og forudse AI-adfærd.
- Forbedring af den brede offentligheds forståelse af AI-teknologier.
- Implementering af robuste meddelelses- og samtykkeprotokoller for at sikre individuel autonomi, når man interagerer med AI-systemer.
- Undgå indsamling og distribution af personlige data uden udtrykkeligt og informeret samtykke fra de berørte personer, med respekt for deres privatliv og uafhængighed.

## TIPS OG ANBEFALINGER TIL UNDERVISERE

Diskuter de mulige autonomiproblemer med AI med eleverne. F.eks. giver Facebook brugerne værktøjer til at udøve kontrol over deres indhold, men kan algoritmens kompleksitet og uigennemsigtighed undertiden underminere ægte brugerautonomi?

## Retfærdighed

Retfærdighed er et mangesidet princip, der søger at bruge AI som et redskab til at rette op på tidligere uretfærdigheder, til at sikre, at fordelene ved AI fordeles retfærdigt, og til at beskytte mod fremkomsten af nye skader eller samfundsmæssige omvæltninger (*disruptions*) (Floridi et al., 2018).

At opretholde retfærdighed i AI, som diskuteret af Jobin et al. (2019), omfatter flere proaktive tiltag:

- Etablering af tekniske standarder og normer.
- Øge gennemsigtigheden af rettigheder og regler.
- Implementering af test, overvågning og revision, især i databeskyttelsesenheder.
- Styrkelse af retsstatsprincippet, herunder appelmuligheder og retsmæssige tiltag.
- Der gennemføres systemiske ændringer, f.eks. regeringstilsyn, fleksible teams og et inkluderende civilsamfund med fokus på en retfærdig fordeling af goderne.

## TIPS OG ANBEFALINGER TIL UNDERVISERE

Se f.eks. tilfældet med Microsofts chatbot Tay: I 2016 afslørede Microsofts racistiske chatbot farerne ved onlinesamtaler - IEEE Spectrum. Med henvisning til retfærdighed prioriterede algoritmen uretfærdigt sensationelt indhold frem for en meningsfuld og konstruktiv diskurs, hvilket førte til en ubalanceret repræsentation af synspunkter og stemmer på platformen og skabte samfundsmæssige omvæltninger.

## Forklarlighed

Forklarlighed er afgørende for at opbygge og bevare brugernes tillid til AI-systemer. Det betyder, at processerne skal være gennemsigtige, at mulighederne og formålet med systemer med kunstig intelligens skal kommunikeres åbent, og at beslutningerne skal forklares til dem, der er direkte og indirekte berørt (High-Level Expert Group on AI (AI HLEG), 2019). Floridi et al. (2018) mener også, at forklarlighed er central i anvendelsen af andre etiske principper.

For at øge gennemsigtigheden i AI-systemer opregner Jobin et al. (2019) flere tilgange:

- AI-udviklere og -brugere opfordres til at dele mere information om deres systemer og afmystificere processerne bag AI-beslutningstagning.

- Forklaringer af AI-operationer og -beslutninger på måder, der er lette at forstå for ikke-eksperter, hvilket sikrer, at AI'ens operationer kan revideres af mennesker, der ikke nødvendigvis er eksperter på området.
- Brugen af audits som en måde, hvorpå man sikrer gennemsigtighed i AI, hvilket kan hjælpe med at identificere og håndtere potentielle bias eller etiske problemer.
- Etablering af mekanismer, der overvåger AI-præstationer, inddragelse af interessenter og offentlighed for at indsamle forskellige perspektiver og mekanismer til at støtte whistleblowing.

## TIPS OG ANBEFALINGER TIL UNDERVISERE

Spørg de studerende, om de føler sig tilstrækkeligt informeret om de beslutninger, der træffes af de AI-løsninger, de bruger. For eksempel manglede Amazons AI-rekrutteringsværktøj gennemsigtighed i sin beslutningsproces. Ansøgere og ansættende ledere fik ikke klare forklaringer på, hvorfor visse ansøgere blev afvist, hvilket gjorde det vanskeligt at udfordre eller tage fat på de underliggende fordomme i systemet.

BEMÆRK! I slutningen af afsnittet om principper skal du bruge det medfølgende casestudie som gruppearbejde.



## 5. Etiske rammer, retningslinjer og værktøjer

Når de etiske principper er blevet klarlagt, skal de omdannes til handlingsorienterede rammer og strategier for at styre AI-baserede innovationer samt den praktiske implementering af AI-etik.

Ifølge Ayling & Chapman (2022) omsætter etiske rammer de etiske principper til praktiske foranstaltninger. Denne proces indebærer at skabe retningslinjer og procedurer, som udviklere og brugere af AI-teknologier kan anvende, og dermed indarbejde etiske overvejelser i hvert trin fra design til anvendelse i den virkelige verden.

Prem (2023) uddyber dette ved at sige, at mange etiske AI-rammer aktivt forsøger at forudse og håndtere potentielle etiske problemer og risici. Ved proaktivt at identificere disse udfordringer kan rammerne være vejledende for, når udviklere skaber AI-løsninger. Denne proaktive tilgang er afgørende på et område, der udvikler sig så hurtigt, som kunstig intelligens gør, hvor nye etiske overvejelser kan dukke op, efterhånden som teknologien udvikler sig og bliver mere integreret i forskellige aspekter af hverdagslivet.

En etisk ramme giver en omfattende plan for etablering af lovgivningsmæssige og operationelle standarder, der harmonerer med etiske normer. Som beskrevet af Floridi & Cowls (2019) er en sådan ramme en integreret del af udformningen af lovgivning, formulering af regler, fastsættelse af tekniske standarder og identifikation af best practice, der kan anvendes på tværs af en række forskellige brancher og geografiske regioner.

Prem (2023, s. 701) definerer de væsentlige komponenter i etiske rammer som følger:

- Grundlæggende **begreber** relateret til diskussionen af etiske aspekter
- Etiske **principper** (f.eks. værdier)
- **Omtanke** for, hvordan brugen og udviklingen af systemer med kunstig intelligens overholder etiske principper
- **Løsninger** på problemerne (f.eks. strategier, regler og retningslinjer)

Værktøjer og retningslinjer for anvendelse af etiske principper i forbindelse med design, implementering og udrulning er meget nødvendige (Zhou et al., 2020). Etiske værktøjer og retningslinjer er vigtige redskaber til at styre AI-innovation i en retning, der er i overensstemmelse med etiske normer og værdier. Etiske værktøjer og retningslinjer ses som vigtige for at bygge bro over kløften mellem abstrakte etiske principper og deres konkrete integration i AI-praksis.

Mens vigtigheden af værktøjer og retningslinjer er klar, er det en udfordring at omsætte principperne til konkrete værktøjer, og mange eksisterende rammer mangler en anvendelseskontekst (Prem, 2023, s. 702). Årsagen til den store kløft mellem principper og praksis skyldes manglende standardisering, foranderlighed, kompleksitet og den subjektive forståelse af de ret abstrakte principper.



Den seneste tids hurtige udvikling af AI har ført til en betydelig vækst i både antallet og mangfoldigheden af tilgængelige værktøjer, og der sker hele tiden ny udvikling. Morley et. al (2019) og Prem (2023) har gennemgået hundredvis af forskellige værktøjer til håndtering af de etiske udfordringer i AI-løsninger. Disse værktøjer er inddelt efter AI-løsningens designfaser og efter forskellige etiske principper. Men mange værktøjer opfattes som vanskelige at bruge eller kræver en betydelig indsats fra brugerne.

Listerne, som er udarbejdet af Morley et. al (2019) og Prem (2023), viser tydeligt, at tilgængeligheden af værktøjer og metoder er ulige fordelt mellem de forskellige principper og designstadier. For eksempel er der mange værktøjer til at måle løsningens nytteværdi i de tidlige udviklingsstadier, mens værktøjer til forklarlighed for det meste er placeret i de senere stadier af udviklingsprocessen. Tjek værktøjerne via dette link: Typologi for anvendt AI-etik - Google Docs.

De anvendte metoder varierer også afhængigt af udviklingsfasen. *Frameworks*, biblioteker og algoritmer spiller ofte en større rolle i *ex ante*-fasen, især i løsningens designfase. Nogle frameworks når dog både *ex ante*- og *ex post*-fasen (UNESCO, 2023b, s. 5). I testfasen er det igen overvejende audits og metrikker, der

præsenteres, mens der i post ante-fasen hovedsageligt bruges *declarations*, *labels* og licenser som metoder (Prem, 2023, s. 709).

Undersøgelsen viste også, at kun få værktøjer adresserede løsningens indvirkning på et individ eller på et samfund som helhed. Det kan skyldes, at det er en udfordring at ændre den komplekse menneskelige adfærd til enkle og generelle designværktøjer. Det er dog vigtigt at tage højde for indvirkningen på mennesker og samfund for at opnå bred accept af og social præference for AI-løsningerne (Morley et al., 2020, s. 2156).

## 6. Fra principper til praksis – rammer for troværdig AI

Selvom der er enighed om, at AI skal være etisk, fortsætter debatterne om den nøjagtige definition af "etisk AI" og de etiske forpligtelser og tekniske kriterier, der er nødvendige for at implementere den (Leijnen et al., 2020).

Europa-Kommissionen har lanceret en omfattende strategi for fremme af kunstig intelligens (AI) i Europa, der lægger stor vægt på at sikre, at den AI, der udvikles og anvendes, ikke kun er avanceret fra et teknologisk synspunkt, men også overholder etiske standarder og er sikker.

For effektivt at implementere denne strategi og dens underliggende principper oprettede Europa-Kommissionen en *High-Level Expert Group on Artificial Intelligence* (AI HLEG), som er en ekspertgruppe på højt niveau om kunstig intelligens. Gruppen fik til opgave at udvikle etiske retningslinjer for kunstig intelligens og anbefalinger til politikker og investeringer, som kan være vejledende for etisk udvikling, implementering og brug af kunstig intelligens i hele Europa og fremme innovation, samtidig med at grundlæggende værdier og rettigheder beskyttes [High-Level Expert Group on AI (AI HLEG), 2019].



Rammen for troværdig AI er designet til at sikre, at AI-teknologier udvikles og implementeres på en måde, der er etisk forsvarlig, respekterer grundlæggende rettigheder og sikrer robusthed og troværdighed. Rammen for troværdig AI består af tre dele, der hver især er designet til at være vejledende for udvikling og anvendelse af AI på en etisk og troværdig måde [High-Level Expert Group on AI (AI HLEG), 2019].

### Grundlaget for troværdig AI

Denne del fastlægger de centrale principper, der ligger til grund for troværdig AI, med en tilgang baseret på grundlæggende rettigheder. Den beskriver de etiske principper, der er afgørende for at sikre, at AI-systemer udvikles og anvendes på en måde, der er både etisk og robust.

Etiske principper udgør det fundament, som troværdig AI skal bygges på, og som sikrer overholdelse af etiske standarder. Disse etiske principper er (i) respekt for menneskelig autonomi, (ii) forebyggelse af skadelige virkninger, (iii) retfærdighed og (iv) forklarlighed.

### Realisering af troværdig AI

Med udgangspunkt i de skitserede etiske principper omsætter den anden del af rammen disse principper til syv centrale krav, som AI-systemer forventes at leve op til i hele deres livscyklus. Denne omdannelse af etiske principper til praktiske krav sikrer, at den grundlæggende etik ikke bare er teoretiske idealer, men er aktivt integreret i alle faser af et AI-systems udvikling og drift.

Listen over krav omfatter: Menneskelig udførelse og kontrol: grundlæggende rettigheder, menneskelig udførelse og kontrol; Teknisk robusthed og sikkerhed: modstandsdygtighed over for angreb, sikkerhed, *fallback*-plan og generel sikkerhed, nøjagtighed, pålidelighed og reproducerbarhed; Beskyttelse af privatlivets fred og datastyring: respekt for privatlivets fred, datakvalitet og dataintegritet samt adgang til data; Gennemsigtighed: sporbarhed, forklarlighed og kommunikation; Diversitet, ikke-diskrimination og retfærdighed: undgåelse af urimelig bias, tilgængelighed og universelt design samt inddragelse af interessenter; Miljø- og samfundsmæssig velfærd: bæredygtighed og miljøvenlighed, social indvirkning, samfund og demokrati; Ansvarlighed: reviderbarhed, minimering og rapportering af negative virkninger, afvejninger og klageadgang.

### Vurdering af troværdig AI

Den tredje del af rammen giver en konkret og omfattende liste til vurdering af troværdig AI, der er designet til at operationalisere de førnævnte krav.



Værktøjet er kendt som "Assessment List for Trustworthy AI (ALTAI)" og indeholder en omfattende og dynamisk tjekliste, der fungerer som en køreplan for udviklere og dem, der implementerer AI, så de kan integrere disse principper i deres AI-applikationer. ALTAI letter denne proces ved at tilbyde et sæt konkrete trin til selvevaluering og derved sikre, at AI-teknologier udvikles på en måde, der maksimerer brugernes fordele og samtidig minimerer eksponeringen for unødvendige risici (High-Level Expert Group on AI (AI HLEG), 2020).

Denne ikke-udtømmende tjekliste fungerer som en praktisk vejledning for AI-praktikere og tilbyder specifik vejledning i, hvordan man implementerer disse krav effektivt. Det er vigtigt, at denne vurdering er fleksibel, så den kan tilpasses den specifikke anvendelse af hvert enkelt AI-system og dermed sikre relevans og effektivitet i forskellige sammenhænge.

## 7. Fra principper til lov: EU'S AI-FORORDNING

AI-forordningen vil regulere anvendelsen af kunstig intelligens i EU, hvilket markerer den første omfattende AI-lovgivning på verdensplan (European Parliament, 2023). Formålet med AI-forordningen er at forbedre det indre markeds funktion og fremme udbredelsen af menneskecentreret og troværdig AI, samtidig med at der sikres et højt niveau af beskyttelse af sundhed, sikkerhed,

grundlæggende rettigheder, herunder demokrati, retsstatsprincippet og miljøbeskyttelse mod skadelige virkninger af systemer med kunstig intelligens i Unionen og støtte innovation (Future of Life Institute, 2024). Det forudses, at loven vil blive indført i 2026 (Hillard & Gulley, 2023).

EU-Kommissionen har organiseret AI-regler inden for en risikobaseret ramme og skabt et system af risikokategorier. AI-applikationer vil blive reguleret i henhold til det risikoniveau, de udgør (European Commission, 2023). Den risikobaserede tilgang, der er beskrevet i EU's AI-forordning, kategoriserer AI-systemer i fire niveauer baseret på deres potentielle indvirkning: uacceptabel risiko, høj risiko, begrænset risiko og lav eller minimal risiko (European Commission, 2023; Hillard & Gulley, 2023).

### Minimal risiko

Størstedelen af AI-systemerne falder i denne kategori og udgør en minimal eller ingen risiko for borgernes rettigheder eller sikkerhed. Eksempler er AI-aktiverede anbefalingssystemer og spamfiltre. Virksomheder, der tilbyder sådanne systemer, er ikke forpligtet til at overholde strenge krav, men kan frivilligt forpligte sig til yderligere adfærdskodekser for at øge gennemsigtigheden og ansvarligheden.

### Begrænset risiko

Denne kategori vedrører AI-systemer, der kun udgør en begrænset risiko for brugernes rettigheder og sikkerhed. Eksempler omfatter chatbots og deep fakes. Selvom de ikke er underlagt strenge lovkrav, har udbyderne pligt til at sikre, at brugerne er opmærksomme på, hvornår de interagerer med AI-systemer og skal mærke AI-genereret indhold i overensstemmelse hermed. Dette sikrer gennemsigtighed og bevidsthed om brugen af AI-teknologier.



### Høj risiko

AI-systemer, der er identificeret som højrisiko, skal overholde strenge krav for at mindske potentielle risici. Disse krav omfatter solide risikoreduktionssystemer, datasæt af høj kvalitet, aktivitetslogning, detaljeret dokumentation, klar brugerinformation, menneskeligt tilsyn og stærke cybersikkerhedsforanstaltninger. Regulatoriske sandkasser, dvs. en slags test-miljøer, etableres for at fremme ansvarlig innovation og lette udviklingen af AI-systemer, der overholder reglerne. Eksempler på AI-systemer med høj risiko omfatter kritisk infrastruktur, medicinsk udstyr, adgangssystemer til uddannelsesinstitutioner og visse anvendelser inden for retshåndhævelse og grænsekontrol.

### Uacceptabel risiko

AI-systemer, der udgør en klar trussel mod grundlæggende rettigheder, er forbudt. Det omfatter systemer, der manipulerer menneskelig adfærd for at underminere menneskets frie vilje, f.eks. legetøj, der tilskynder mindreårige til farlig adfærd, eller systemer, der muliggør "social scoring" for regeringer eller virksomheder. *Predictive policing*-applikationer (dvs. brug af big data til at

forudsige og forebygge kriminalitet) samt anvendelse af biometriske systemer, som f.eks. følelsesgenkendelse på arbejdspladser, er også forbudt.

## TIPS OG ANBEFALINGER TIL UNDERVISERE

1) Diskuter med eleverne, om de mener, at EU-lovgivningen er strengere end andre love, og hvad den vil medføre. Stof til eftertanke, f.eks. fra denne artikel: https://news.bloomberglaw.com/artificial-intelligence/eu-embraces-new-ai-rules-despite-doubts-it-got-the-right-balance
2) Hvis man har den seneste diskussion om TikTok in mente, og hvor skadeligt det er, hvilken risikokategori ville det så falde ind under?

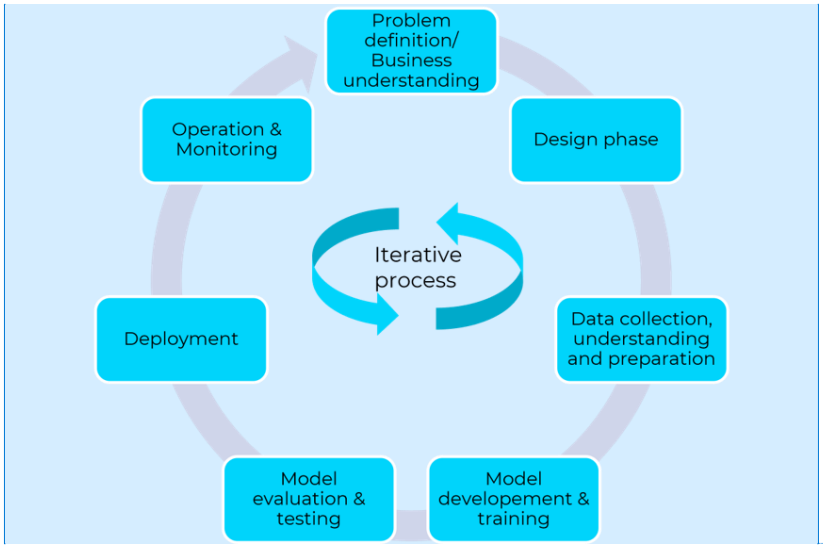
## 8. Livscyklus for AI-udvikling

Som det fremgår, bliver specifikke etiske principper mere relevante i visse faser af udviklingen af AI-systemer, og forskellige værktøjer anvendes i overensstemmelse hermed. Det er afgørende at forstå udviklingsprocessen ud fra et bredt perspektiv. (Prem, 2023, s. 703) I dette kursus anvender vi udtrykket "AI-udviklingens livscyklus" for at sammenfatte dette vidtgående syn på udviklingsstadierne.



Den hurtigt udviklende branche inden for AI-løsninger oplever nogle gange projektfejl på grund af utilstrækkelige projektstyringsprocesser. En veldefineret AI-livscyklus, der tager højde for alle udviklingstrin, kan forbedre AI-løsningernes succesrate. Der findes forskellige versioner af AI-livscyklusbeskrivelser, men vores tilgang sammenfatter de modeller, der er foreslået af Morley et al. (2020), Prem (2023) og CRISP-modellen fra Data Science Process Alliance (Hotz, 2018) som præsenteret af Rochel & Evequoz (2021).

Livscyklussen for AI-udvikling beskriver den rejse, et AI-system tager fra den første idé undfanges til den endelige implementering og vedligeholdelse. Hver fase er afgørende for at sikre et vellykket resultat. Designlivscyklussen er iterativ, hvilket betyder, at det kan være nødvendigt at genbesøge tidligere faser, hvis visse variabler ikke fungerer som forventet. Selvom etiske overvejelser kan variere på tværs af forskellige faser, er det vigtigt at integrere disse overvejelser i alle faser af processen. (Saltz, 2023)




Comentado [KM3]: Needs to be translated! Terminology must be the same as in the text.

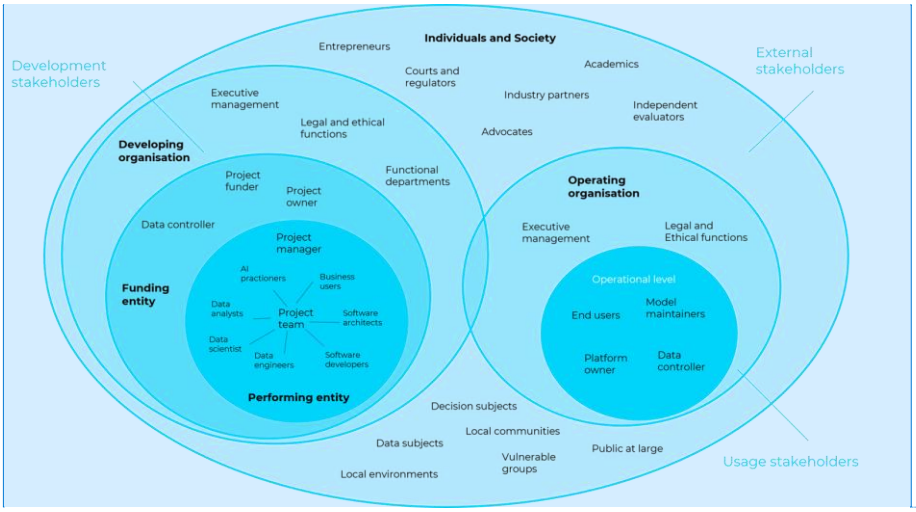
Billede: Forfattere af dokumentet. Modificeret som en kombination af de forskellige modeller, der er beskrevet i artikler skrevet af Data Science Process Alliance (2023), Morley et. al. (2019), Prehm (2022) og Rochel et. al. (2020).

## 9. Interessenternes engagement i AI-udviklingens livscyklus

Inddragelse af en bred vifte af interessenter i AI-udviklingsprocessen er afgørende for en grundig evaluering af de etiske konsekvenser af et AI-projekt fra flere perspektiver. At inddrage interessenter, herunder dem, der er direkte påvirket af AI-løsningen, er afgørende for at opbygge pålidelige systemer (Morley et al., 2020) og for en omfattende vurdering af projektets virkninger (UNESCO, 2023a, s. 15). Dette kapitel giver et overblik over de forskellige interessenter, der er involveret i AI-udviklingsprojekter, og beskriver deres roller i forskellige faser af AI-udviklingens livscyklus.

Interessenterne opdeles i udviklingsinteressenter, brugsinteressenter og eksterne interessenter. Udviklings- og brugsinteressenter er aktivt involverede interessenter, som f.eks. beslutningstagerne, der har magten til at træffe beslutninger, og designerne, som har ekspertisen. De eksterne interessenter er ikke involveret i AI-projektet, men kan påvirke eller blive påvirket af det eller blive konsulteret. (Miller, 2022)

Dernæst undersøges de forskellige interessenters nøgleroller i udviklingen af kunstig intelligens, og de inddeles i tre kategorier: udvikling, brug og eksterne interessenter. Desuden undersøges hver gruppes specifikke ansvar og bidrag i forskellige faser af AI-udviklingens livscyklus. Det er vigtigt at forstå disse roller for at kunne navigere i AI-udviklingens kompleksitet og sikre, at projektet bliver en succes.




Billede: G.J. Miller (2022)

## Udviklingsinteressenter

Inden for AI-projektudvikling opdeler Miller (2022) interessenter i tre hovedkategorier: organisationen, den finansierende enhed og den udførende enhed, som hver især spiller en særskilt rolle i projektets livscyklus. Organisationen drager fordel af resultaterne af AI-projektet og udnytter fremskridtene og innovationerne til sin vækst og succes. Finansieringsenheden, som er afgørende for den økonomiske støtte, finansierer ikke kun, men styrer også projektet og sikrer, at investeringerne er i overensstemmelse med projektets mål og styringsstandarder. Den udførende enhed er afgørende, idet den tager ansvaret for at generere projektets output og udforme AI-systemet, der repræsenterer projektets tekniske udførelse (Miller, 2022).

**Comentado [KM4]:** Should be translated.

Inden for organisationen har rollerne et spænd fra udviklingsorganisationen, som sponsorerer og definerer projektets omfang, til den øverste ledelse, som fører tilsyn med projektets forløb. De driftsafdelingerne yder vigtig støtte, mens juridiske og etiske funktioner sikrer, at projektet er i overensstemmelse med lovmæssige standarder og etiske normer. Projektledelsesgruppen spiller en støttende rolle og hjælper projektlederne med administrative opgaver for at strømline projektudførelsen (Zwikael & Meredith, 2018).

Finansieringsenheden omfatter roller som projektfinansiereren, der leverer finansielle ressourcer, og projektejeren, der giver strategisk retning og potentielt er formand for styregruppen og sikrer, at projektet er i overensstemmelse med de strategiske mål. I scenarier, hvor den finansierende enhed ikke er direkte involveret, uddelegeres ansvaret for at sikre løbende tilsyn. Den dataansvarlige, en rolle, der ofte er forbundet med den finansierende enhed, administrerer dataanvendelsen og sørger for at overholde strenge lovgivningsmæssige og juridiske rammer for at undgå potentielle sanktioner eller bøder (Miller, 2022).

Kernen i den udførende enhed er projektlederne og projektteamet, der er ansvarlige for at levere projektets output i henhold til den godkendte plan. Dette samarbejde sikrer, at projektet skrider frem fra idé til virkelighed, hvor hvert teammedlem bidrager til udviklingen og den vellykkede implementering af AI-systemet (Zwikael & Meredith, 2018).



Denne strukturerede organisering af interessenter og deres roller understreger den samarbejdsorienterede og mangesidede tilgang, der er nødvendig for en vellykket udvikling og implementering af AI-projekter, og fremhæver vigtigheden af klare roller, ansvarsområder og overholdelse af juridiske og etiske standarder.

### Brugsinteressenter

I forbindelse med AI-projekter spiller brugerinteressenter en afgørende rolle, især på det organisatoriske niveau, hvor driftsorganisationen fungerer som en kundeinteressent, der anskaffer AI-løsningen. I lighed med udviklingsorganisationen omfatter driftsorganisationen roller som ledelse og juridiske og etiske funktioner, hvilket fremhæver de parallelle strukturer i både udviklings- og driftsfasen (Miller, 2022).

På det operationelle niveau involverer brugsfasen en række nøgleaktører, herunder dataforvaltere, slutbrugere, platformsejere og dem der skal vedligeholde systemet. Slutbrugere, uanset om de er enkeltpersoner eller grupper, interagerer med AI-systemet enten gennem arbejds- eller servicekontrakter med driftsorganisationen eller som forbrugere, der engagerer sig direkte i systemets funktionaliteter (Miller, 2022).

De, der skal vedligeholde systemet, påtager sig den vigtige opgave med løbende at styre, opdatere og forbedre de maskinlæringsmodeller, der driver AI-systemet, og sikrer dets effektivitet. Dataansvarlige har til opgave at definere målene og

metoderne for behandling af persondata i systemet og sikre, at databeskyttelseslovgivningen overholdes. Endelig har platformsejere ansvaret for AI-systemets overordnede infrastruktur og dets implementering og opretholder de grundlæggende forhold, som AI fungerer på (Miller, 2022).

### Eksterne interessenter

Miller (2022) kategoriserer eksterne interessenter i AI-projekter i tre hovedundergrupper: enkeltpersoner, samfundet og repræsentanter, som hver især har forskellige roller og bidrag til projektet.

Undergruppen 'enkeltpersoner' omfatter datasubjekter, beslutningssubjekter og medarbejdere, der interagerer med udviklings- eller driftsorganisationen gennem aftaler som service- eller ansættelseskontrakter. Disse interessenter omfatter personer, der er direkte involveret i eller påvirket af projektet. Disse interessenter spiller afgørende roller som datasubjekter, beslutningssubjekter og medarbejdere i projektet (Miller, 2022).

Undergruppen 'samfundet' omfatter den bredere indvirkning på den generelle befolkning, herunder lokalsamfund, miljøet og sårbare grupper som mennesker med handicap, minoriteter, mindreårige og offentligheden som helhed. Dette segment påvirkes direkte af udviklingen og brugen af AI-systemet, hvilket understreger AI-projekternes samfundsmæssige rækkevidde (Miller, 2022).



'Repræsentanter', som udgør den tredje undergruppe, adskiller sig fra formelle repræsentanter. Mens repræsentanter omfatter grupper og organisationer som iværksættere og uafhængige evaluatorer (rapportører, revisorer eller anmeldere), der måske ikke er direkte berørt af AI-systemet, men spiller en rolle i dets udvikling eller evaluering, repræsenterer formelle repræsentanter som domstole, love, tilsynsmyndigheder og fagforeninger juridisk og lovgivningsmæssigt tilsyn og styring inden for projektet (Miller, 2022; Rodrigues, 2020).

### Inddragelse af interessenter i udviklingen af AI-løsninger

Samskabelse refererer til samarbejde med interessenter, herunder slutbrugere eller kunder, gennem hele designprocessen. Denne tilgang letter udvekslingen af indsigt mellem deltagere med forskellige roller og ekspertise (Interaction Design Foundation - IxDF, 2021; Sanin, 2020).

Samskabelse kan udføres via faciliterede workshops, der omfatter en række forskellige aktiviteter, f.eks. *ice-breakers*, rollespil, kortlægning af forløb, brainstorming og prototyping. Ved at anvende samskabelse får designerne mulighed for at få en dyb forståelse af kundernes eller slutbrugernes behov og ønsker og dermed skabe løsninger, der tager hensyn til forskellige brugerperspektiver. Denne inkluderende metode sikrer, at de udviklede tjenester er godt afstemt med de faktiske krav fra dem, de sigter mod at betjene (Interaction Design Foundation - IxDF, 2021; Sanin, 2020).

## 10. Livscyklus for AI-udvikling: Problemformulering og forretningsforståelse

### TIPS OG ANBEFALINGER TIL UNDERVISERE

I dette afsnit forklares – i detaljer – livscyklussen for udviklingen af en AI-løsning, dens etiske konsekvenser og de implicerede interessenter. For at fremme en livlig diskussion kan du bruge et eksempel til at gennemgå hele livscyklussen, f.eks. det AI-baserede anbefalingssystem om, hvad man kunne købe, se dette link: https://www.datascience-pm.com/ai-lifecycle/#An_Example_Ai_Project_Life_Cycle

Som alle andre projekter bør AI-livscyklussen også starte med en problemformulering. En veldefineret problemformulering giver en dyb forståelse af en kundes behov og dermed også en retning for projektet. På dette tidspunkt bør der også gennemføres en forundersøgelse for at fastslå, om AI er den mest hensigtsmæssige løsning på problemet. Det er også nødvendigt at inddrage forskning og interessenter for at forstå problemet til bunds (De Silva & Alahakoon, 2022, s. 4-5).



### Etiske principper i fasen for problemformulering og forretningsforståelse

I den indledende fase af et projekt er det afgørende at overveje de væsentligste etiske aspekter, da de er udgangspunktet for projektets videre rejse og den efterfølgende samfundsmæssige indvirkning. Særligt i denne fase anses principperne om gavnlighed og ikke-skadelighed (*non-maleficence*) for at være altafgørende (Prem, 2023, s. 703).

Med hensyn til gavnlighed er fokus på at vurdere, om AI-systemet aktivt fremmer individers og samfundets ve og vel. Systemets formål skal være gennemsigtige og sigte mod at give reelle fordele, samtidig med at der værnes om grundlæggende rettigheder, og der tages hensyn til miljømæssig bæredygtighed. Det er vigtigt at inddrage en række forskellige interessenter og sikre, at holdningerne hos dem, der er berørt af AI-løsningen, bliver hørt og integreret i løsningen (De Silva & Alahakoon, 2022, s. 5; Morley et al., 2020, s. 2151).

Princippet om ikke-skadelighed indebærer at sikre, at AI-systemet ikke påfører enkeltpersoner eller samfund skade. Det kræver en proaktiv tilgang at identificere og afbøde potentielle skadelige virkninger og årvågent beskytte det fysiske og mentale velbefindende hos de mennesker, der påvirkes af systemet (De Silva & Alahakoon, 2022, s. 5; Morley et al., 2020, s. 2151). Generelle

sikkerhedshensyn skal adresseres proaktivt og indbygges i systemets design (UNESCO, 2023b, s. 18).

Ud fra et juridisk synspunkt er det vigtigt at undersøge AI-systemets generelle konsekvenser for institutioner, demokrati og samfundsstrukturer. Det er vigtigt at overveje, om projektet utilsigtet kan skabe bias, favorisere visse grupper frem for andre eller gribe ind i individets autonomi. Hvert af disse aspekter kræver nøje overvejelser for at sikre, at AI-systemet bidrager til et fair og retfærdigt samfund (Morley et al., 2020, s. 2151).

### Interessenter i fasen for problemformulering og forretningsforståelse

I denne fase er det vigtigt som minimum at inddrage slutbrugerne, som er de primære brugere af AI-løsningen. Derudover bør AI-udviklere inddrages tidligt for at få en grundig forståelse af det problem, de får til opgave at løse. Etikere eller samfundsforskere bør også inddrages for at vurdere de potentielle menneskelige og samfundsmæssige konsekvenser af løsningerne. Domæneeksperter er også vigtige at inddrage for at øge viden om den branche, hvor AI skal anvendes (De Silva & Alahakoon, 2022).

## 11. Livscyklus for AI-udvikling: Designfasen



I designfasen videreudvikles det grundlæggende arbejde, som blev udført i problemformuleringsfasen, til en udførlig plan for implementeringen af AI-systemet. Denne fase omfatter definition af projektmål og ønskede resultater, udvikling af en projektplan med tidslinjer og budgetovervejelser samt en forståelse af interessenternes krav (De Silva & Alahakoon, 2022, s. 4).

De vigtigste aktiviteter omfatter udarbejdelse af en kravspecifikation, der specificerer AI-systemets funktioner og mål for projektets succes. Derudover finpudses strategierne for datamining, herunder f.eks. beslutningerne om brug af præ-trænede modeller. I hele denne fase bør etiske overvejelser inddrages i alle aspekter af systemets design (De Silva & Alahakoon, 2022, s. 3).

### Etiske principper i designfasen

I AI-løsningens designfase er det afgørende at anvende proaktive "hvad nu hvis"-strategier for at forudse og afbøde potentielle udfordringer i hele AI-løsningens livscyklus. Ved at stille kritiske spørgsmål som: "Hvad nu, hvis de anvendte data er biased?" og finde de nødvendige mekanismer til at opdage og håndtere retsstridig adfærd, kan udviklere forebygge disse udfordringer. Det er vigtigt at etablere konkrete foranstaltninger for at udstyre udviklere med viden, der kan afværge bias relateret til f.eks. race, hudfarve, afstamning, køn, alder, sprog, religion, politisk overbevisning, national eller etnisk oprindelse, social baggrund,

økonomisk status, opvækstbetingelser eller handicap. Desuden er det vigtigt at etablere robuste kommunikationskanaler, rapporteringsprotokoller og foranstaltninger, der er hurtige til at reagere på eventuel bias og de justeringer, de måtte kræve (Morley et al., 2020, s. 2151, 2155; UNESCO, 2023b, s. 13, 17).

Ud over disse forholdsregler bør designkravene klart stipulere metoder til at øge gennemsigtigheden i processen og dermed øge forklarligheden. Denne tilgang understreger vigtigheden af at gøre AI-systemets operationer forståelige og ansvarlige (Morley et al., 2020, s. 2151, 2155; UNESCO, 2023b, s. 13, 17).

Prem (2023, s. 703) understreger desuden værdien af at inddrage interessenter i denne fase og den vigtige rolle, som menneskers overvågning af systemet spiller. Denne inddragelse sikrer, at en bred vifte af perspektiver bidrager til udviklingen af AI, hvilket fremmer en mere retfærdig og informeret designproces.

### Interessenter i designfasen

I denne fase er det vigtigt at involvere softwarearkitekter, da de spiller en afgørende rolle i at lægge det tekniske fundament for AI-løsningen og overveje både aktuelle krav og fremtidig skalerbarhed (Peter, 2024). UX-designere er afgørende for at skabe brugergrænseflader, der fremmer brugernes accept og tilfredshed med systemet, og som sikrer, at brugerne kan interagere problemfrit med AI-systemet (White, 2023). AI-praktikere skal samarbejde tæt med UX-designere for at integrere AI-komponenter smidigt i brugergrænsefladen. Etik-eksperter kan adressere etiske spørgsmål i forbindelse med AI-udvikling. Der er brug for juridiske og etiske instanser for at reducere de juridiske risici, der er forbundet med AI-projekter.

Ifølge De Silva og Alahakoon (2022) anbefales det at bruge en participatorisk designtilgang, der involverer de mennesker og det samfund, der vil blive berørt af AI-modellen og dens beslutninger, og som fremmer gennemsigtighed, inklusion og tilpasning til samfundsmæssige værdier i hele projektets livscyklus.

## 12. Livscyklus for AI-udvikling: Dataindsamling, -forståelse og -forberedelse

Når man har sat sig et mål og lagt en strategi, er næste fase at indsamle og forberede relevante data. Dette omfatter:

- At indsamle data
- At definere og kategorisere data
- At gennemføre en eksplorativ analyse
- At validere dataenes relevans
- At udvælge vigtig information

- At vaske data, rekonstruere, integrere og formatere dataene, så de passer til projektets formål

Det er afgørende at håndtere og kompensere for eventuelt manglende data-værdier. Det er også vigtigt at forstå antagelserne bag de eksisterende data og sikre klarhed over, hvordan datasættene bruges i beslutningsprocesserne. Opretholdelse af datakvalitet er altafgørende, især hvis dataene er ufuldstændige eller tvetydige (Rochel & Evéquoz, 2021, s. 614-617).

Selv om denne fase er særdeles tidskrævende, er den også helt essentiel. Troværdigheden og ydeevnen af den færdige AI-løsning hænger direkte sammen med kvaliteten og nøjagtigheden af de underliggende data (Saltz, 2023).

## Etiske principper i dataindsamlings-, forståelses- og forberedelsesfasen

Dataindsamlingsfasen er forbundet med risici, især hvad angår princippet om ikke-skadelighed. Det er afgørende at sikre databeskyttelse og fortrolighed i denne fase. AI-systemer skal være konsekvente med hensyn til databeskyttelse, lige fra starten og gennem hele løsningens livscyklus. Juridiske bestemmelser, som f.eks. global databeskyttelseslovgivning, regulerer typisk spørgsmål om beskyttelse af privatlivets fred (De Silva & Alahakoon, 2022, s. 5).

Bias i dataene kan føre til overrepræsentation af visse demografiske grupper, så det er vigtigt at undersøge og sikre dataenes kvalitet og integritet. Derudover kan datasæt være udsat for cybertrusler, hvilket kræver robuste sikkerhedsforanstaltninger for at beskytte dem mod potentielle angreb (UNESCO, 2023a, s. 9, 18).



At være gennemsigtig i dataindsamlingsprocesserne, herunder hvilke data der indsamles, til hvilke formål, og hvordan de vil blive brugt i AI-systemet, er også relevant på dette tidspunkt (De Silva & Alahakoon, 2022, s. 6).

## Interessenter i dataindsamlings-, forståelses- og forberedelsesfasen

Det er afgørende at inddrage dataforskere i denne fase, da de spiller en central rolle i at skelne mellem mønstre og tendenser i dataene og omdanne dem til brugbar viden (Miller, 2022). På samme måde er AI-teknikere uundværlige til at vurdere og verificere dataenes integritet (Rochel & Evéquoz, 2021), mens dataanalytikere forbedrer projektet ved at forstå dataenes egenskaber og rådgive om effektive strategier for, hvordan data kan forberedes. Derudover er data-teknikeres arbejde afgørende for forberedelsen og vedligeholdelsen af dataene, hvilket sikrer deres kvalitet (Miller, 2022).

Det er vigtigt at udpege en databeskyttelsesrådgiver (DPO) for at sikre overholdelse af juridiske og etiske standarder for persondata (European Data Protection Supervisor, 2024).

Etikere spiller en vigtig rolle ved at integrere etiske overvejelser i hele datahåndterings- og udvælgelsesprocessen. Ved at inddrage disse forskellige interessenter får AI-projektet tilført omhyggeligt forberedte data, som er fremskaffet på en etisk forsvarlig måde, hvilket skaber et fundament af tillid og troværdighed i AI-løsningen.

## 13. Livscyklus for AI-udvikling: Modeludvikling og træning

I modeludviklings- og træningsfasen er oprettelsen af en AI-model specifikt rettet mod at tackle den udfordring eller det problem, som indtil nu er defineret. Denne fase omfatter omhyggelig udvælgelse, opsætning og den påtænkte brug af algoritmiske værktøjer, hvorfor det er vigtigt at kunne dokumentere og vurdere konsekvenserne af disse valg grundigt, især ved at identificere eventuelle svagheder eller punkter, hvor det er nødvendigt at indgå kompromiser. Fasen er også kendetegnet ved at være iterativ, hvor der vil være flere runder af videreudviklinger for at forbedre modellens nøjagtighed og effektivitet.

Processen med at vælge algoritmiske værktøjer indebærer ofte, at man skal navigere mellem formål, der kan være modstridende. Et godt eksempel er udfordringen med at filtrere spammails, hvor der er behov for at afbalancere målet om at minimere spam i brugernes indbakker med risikoen for fejlagtigt at kategorisere "rigtige" e-mails som spam. At opnå den optimale balance er afgørende for at forbedre modellens effektivitet uden at gå på kompromis med dens troværdighed eller brugernes tillid (Rochel & Evéquoz, 2021, s. 617-618; Saltz, 2023).



### Etiske principper i modeludviklings- og træningsfasen

Selv om tekniske detaljer ofte har forrang i denne fase, giver fasen en god mulighed for at sikre, at den valgte model kan forklares og forstås, samt for at vurdere, om der eksisterer bias i den (Prem, 2023, s. 703). Udviklingsprocessen skal være tydelig og forståelig for alle involverede parter, hvilket kræver grundig dokumentation af datakilder, modelvalg og den logik, der ligger til grund for disse beslutninger (Morley et al., 2020, s. 2151).

Når det er nødvendigt at indgå kompromiser, især fordi det at arbejde effektivt ofte indebærer at indgå kompromiser, er det vigtigt med en systematisk tilgang til at forstå og åbent vurdere kompromisernes virkninger. AI-teknikere skal være parate til at begrunde deres valg og tvinges til at redegøre for og afbøde eventuelle negative konsekvenser for dem, der påvirkes (Morley et al., 2020, s. 2151).

### Interessenter i modeludviklings- og træningsfasen

Forskere inden for AI eller *machine learning* spiller en central rolle i at transformere problemformuleringen til den første AI-modelprototype (De Silva & Alahakoon, 2022), hvilket viser, hvor centrale de er, når projektet går videre til modeludviklings- og træningsfasen. I denne fase er AI-teknikere uundværlige på grund af deres evne til at identificere og anvende de bedst egnede algoritmiske værktøjer, der er specifikt tilpasset projektets formål (Rochel & Evéquoz, 2021). Desuden er dataforskere kernen i denne fase, idet de styrer udviklingen og den løbende forbedring af maskinlæringsmodellerne i retning af større nøjagtighed og effektivitet.

Inddragelse af etiske eksperter er altafgørende for at sikre, at etiske overvejelser indarbejdes i udviklingsprocessen, så man undgår eventuelle negative konsekvenser for involverede brugere af systemet. Derudover er den indsigt og forståelse, som brugere i en virksomhed måtte opnå, uvurderlig, da den er med til at sikre, at de udviklede modeller er i overensstemmelse med forretningsstrategier og opfylder de forventede krav og dermed bygger bro mellem teknisk innovation og praktisk anvendelse i virksomheden.

## 14. Livscyklus for AI-udvikling: Evaluering og afprøvning af modeller



Når AI-modellen er udviklet og trænet, skal dens effektivitet testes og evalueres i forhold til foruddefinerede mål og succeskriterier. I evalueringsfasen måles ikke kun modellens performance, men de underliggende antagelser og udvælgelseskriterier i de anvendte datasæt undersøges også. I tilfælde, hvor modellen ikke lever op til forventningerne, kan der være behov for justeringer, hvad enten det er i modellens parametre, dens overordnede arkitektur eller de datasæt, den bygger på. Her er det afgørende, at der mellem projektlederen og alle relevante interessenter er åben kommunikation af eventuelle fejl og mangler, også i betragtning af den potentielle betydelige indvirkning på de involverede personer. På baggrund af evalueringsresultaterne skal teamet beslutte de efterfølgende tiltag, som kan indebære yderligere iterationer til forbedring af modellen, eller at man fortsætter til implementeringsfasen, hvor det skal sikres, at hvert trin er klart afstemt og aftalt (Hotz, 2018; Rochel & Evéquoz, 2021, s. 619).

### Etiske principper i modelevaluerings- og testfasen

I denne fase er det altafgørende at teste modellen for, om den overholder etiske retningslinjer, og sikre, at den følger de vigtige principper, der blev fastlagt i begyndelsen. Udvikling og finpudsning af audits og metrikker for reviderbarhed er vigtige opgaver hér. Det er vigtigt at få indarbejdet de vigtige etiske principper som retfærdighed, gavnlighed og ikke-skadelighed. Det er afgørende at vurdere modellens modstandsdygtighed over for potentielle sikkerhedstrusler og udarbejde en strategi for uventede udfordringer. Der bør være klare mekanismer

for korrigerende handlinger i tilfælde af modstridende eller uønskede resultater. Det er nødvendigt med en grundig evaluering, der omfatter aspekter som beskyttelse af privatlivets fred, eventuelle sociale konsekvenser, potentielle bias m.m., og en forpligtelse til at minimere og åbent rapportere eventuelle negative konsekvenser (De Silva & Alahakoon, 2022, s. 8; Morley et al., 2020, s. 2151; Prem, 2023, s. 703).

Gennemsigtighed er afgørende, ikke kun med hensyn til modellens tekniske funktion, men også med hensyn til, hvordan beslutninger er truffet i løbet af projektet, hvilket sikrer forklarlighed (De Silva & Alahakoon, 2022, s. 8; Morley et al., 2020, s. 2151). Derudover er det afgørende at forstå, at udviklere kan blive presset til at fremskynde udviklingsprocessen med henblik på en hurtigere implementering, hvilket kan have negative konsekvenser (Rochel & Evéquoz, 2021, s. 618).

### Interessenter i modelevaluerings- og testfasen

I denne fase spiller AI-teknikere en vigtig rolle, ved at de kan fortolke resultaterne fra modelleringen, hvilket giver vigtig viden om resultaterne (Rochel & Evéquoz, 2021).

Kvalitetssikringsteams (QA-teams) er også vigtige, da de kan bekræfte, at modellerne overholder kvalitetsstandarder og er uden fejl eller utilsigtede resultater. Inddragelse af etik-eksperter er nøglen til at garantere en etisk anvendelse af AI-modeller og sikre, at de bruges ansvarligt. Repræsentanter fra potentielle brugergrupper bidrager væsentligt ved at tilpasse AI-modellerne til brugernes forventninger og krav, hvorved brugernes tilfredshed og accept af systemet øges.

Derudover er *Legal and Compliance Teams* uundværlige i forhold til at håndtere juridiske risici og sikre, at AI-modellerne overholder alle gældende love og regler (Moore, 2023).



## 15. Livscyklus for AI-udvikling: Implementering

Efter en vellykket udviklingsfase implementeres AI-løsningen f.eks. i et produktionsmiljø med det formål at løse problemer i den virkelige verden. Omfanget af og metoden til implementering afhænger af løsningens specifikke formål og de systemer eller applikationer, den er designet til. Typisk begynder udrulningen med en pilotfase med en udvalgt gruppe af eksperter og førstegangsbrugere, som kan give deres umiddelbare feedback (De Silva & Alahakoon, 2022, s. 8).

AI-løsningen kan integreres i eksisterende systemer, danne grundlag for en ny applikation eller service, eller dens resultater kan deles via offline-metoder som

f.eks. ledelsesrapporter. Det er afgørende at etablere en feedback-mekanisme til at overvåge dens performance og indvirkning, så man kan foretage løbende forbedringer baseret på, hvordan den bruges i den virkelige verden (Saltz, 2023).

For at håndtere potentielle risici implementeres omfattende risikostyringsstrategier. Der bør indføres drifts- og overvågningsforanstaltninger for at sikre, at AI-løsningen fungerer effektivt efter implementeringen. Der bør også foretages en grundig projektgennemgang til sidst, som resulterer i en detaljeret rapport, der opsummerer projektets resultater og den opnåede læring (De Silva & Alahakoon, 2022, s. 8-9).

## Etiske principper i implementeringsfasen

Morley (2020) fremhæver, at etiske overvejelser ofte får utilstrækkelig opmærksomhed i implementeringsfasen. At forenkle kompleks menneskelig adfærd til værktøjer, der både er forståelige og gennemsigtige, er en stor udfordring. Alligevel er det afgørende at skabe sådanne værktøjer for at øge tilliden til AI-systemer. Derfor fremstår forklarlighed som et vigtigt etisk princip i denne fase.

Desuden er det vigtigt at sikre menneskelig autonomi ved at informere slutbrugerne, når de interagerer med en AI-drevet proces. Denne gennemsigtighed er afgørende for at undgå overdreven afhængighed af AI-systemer. UNESCO (2023b, s. 36) understreger behovet for effektive mekanismer på fire hovedområder: systembevidsthed, robuste revisionsprocedurer, klarhed over, hvordan algoritmerne fungerer, og foranstaltninger til at måle systemernes indvirkning, herunder hvordan man kan indgive en klage eller anke, som også bør omfatte beskyttelse af whistleblowers. Derudover kan den vigtige rolle, som menneskeligt tilsyn spiller, ikke overvurderes, og der er behov for bestemmelser, der tillader menneskelig indgriben og tilsyn med AI-systemer for at sikre, at de anvendes etisk korrekt (Morley et al., 2020, s. 2151).



## Interessenter i implementeringsfasen

Teknikere inden for AI eller machine learning er afgørende for at omdanne prototypen på en AI-model til en fuldt implementeret service eller løsning, der er tilgængelig for interessenter og slutbrugere (De Silva & Alahakoon, 2022).

Cybersikkerhedseksperter bør foretage en risikovurdering for at sikre AI-systemet (Junklewitz et al., 2023, s. 9-12). Slutbrugerne bidrager væsentligt ved at påpege eventuelle problemer eller mangler, som måske ikke har været tydelige under test i kontrollerede miljøer. Derudover er brugernes talspersoner eller repræsentanter medvirkende til at indsamle feedback om den implementerede AI-løsnings anvendelighed og performance. Deres rolle sikrer, at løsningen stemmer overens med brugernes forventninger og krav og muliggør forbedringer, hvor det er nødvendigt.

## 16. Livscyklus for AI-udvikling: Drift og overvågning

Løbende vedligeholdelse, opdateringer og evalueringer er afgørende for et operationelt AI-systems levetid og effektivitet. Der bør konstant blive foretaget performance reviews for at sikre, at systemet lever op til de forventede standarder, mens løbende overvågning af brugerinteraktioner giver værdifuld indsigt i den virkelige verdens anvendelse af systemet, og hvor der er potentielle forbedringsområder. Regelmæssige opdateringer med nye data er afgørende for, at modellen forbliver relevant og præcis. Brugerfeedback er afgørende for den løbende forbedring af AI-modellen, hvilket understreger vigtigheden af iterative forbedringer som en del af vedligeholdelsescyklussen. Dette beviser også, at modellen er et fremskridt snarere end fiasko.

Derudover er det vigtigt at måle systemets investeringsafkast (ROI) og andre vigtige parametre for at vurdere, om det lykkes at nå de foruddefinerede mål (De Silva & Alahakoon, 2022, s. 9; Saltz, 2023).

### Etiske principper i drifts- og overvågningsfasen

I denne fase omfatter overvågningen af modellens performance mere end blot tekniske målinger; den omfatter også overholdelse af etiske standarder. Regelmæssig revurdering af alle principper er afgørende, især i betragtning af mulige ændringer i datasæt og modelstruktur. Det er vigtigt at inddrage interessenter, også efter implementeringen, og deres input er afgørende for at der kan ske løbende forbedringer. Man bør aktivt efterspørge feedback fra brugere, organisationer og det bredere samfund. Der skal være mekanismer på plads til konsekvent at indsamle og indarbejde denne feedback i den løbende forbedring af modellen (High-Level Expert Group on AI (AI HLEG), 2019, p. 19; UNESCO, 2023b, p. 15).

### Interessenter i drifts- og overvågningsfasen

I drifts- og overvågningsfasen er *DevOps* og *IT Operations Teams* afgørende for at overvåge systemets performance, og at systemet er tilgængeligt hele tiden (Immersant Data Solutions, 2023). Uafhængige evaluatorer, som f.eks. fagfolk, der kan vurdere og certificere sikkerheden, kan vurdere sikkerheden i den udviklede AI-løsning (Miller, 2022). Derudover kan ekspertpaneler, styregrupper eller lovgivningsorganer gennemgå projektet med hensyn til både tekniske og etiske aspekter (De Silva & Alahakoon, 2022).

## 17. Konklusion

I denne enhed har vi set nærmere på de etiske principper, rammer og retningslinjer, der er afgørende for udviklingen af AI-løsninger. Vi har gennemgået AI-udviklingens livscyklus, fra den indledende problemformulering til drift og overvågning af AI-løsningen, med stort fokus på integration af relaterede principper og inddragelse af interessenter i alle faser.

Etisk AI-udvikling viser, hvor vigtigt det er at indarbejde etiske overvejelser i hele AI-udviklingens livscyklus. Denne tilgang sikrer, at AI-teknologier ikke kun bidrager positivt til samfundet, men også håndterer potentielle risici såsom bias og bekymring for privatlivets fred effektivt. Desuden blev betydningen af at inddrage et bredt spektrum af interessenter understreget, hvilket illustrerer, at vellykket AI-udvikling kræver mere end blot teknisk ekspertise for at forskellige perspektiver bliver inkluderet. Denne inkluderende tilgang beriger udviklingsprocessen og sikrer, at AI-løsninger er i overensstemmelse med samfundets værdier og opnår dermed bredere accept.

Desuden blev behovet for løbende overvågning og iterativ forbedring af AI-systemer efter udrulningen understreget som afgørende for at opretholde systemernes etiske integritet og sikre, at de forbliver brugbare og relevante. Ved at indtage en proaktiv holdning til innovation kan udviklere identificere og håndtere etiske udfordringer, når de opstår, og dermed navigere i kompleksiteten i AI-udvikling med en forpligtelse til overholdelse af etiske standarder og udføre ansvarlig innovation.



## 18. Referencer

Ayling, J., & Chapman, A. (2022). Putting AI ethics to work: Are the tools fit for purpose? AI and Ethics, 2(3), 405–429. https://doi.org/10.1007/s43681-021-00084-x

Beijing Academy of Artificial Intelligence—Beijing AI Principles. (2019). Datenschutz Und Datensicherheit, 10.

Beijing Artificial Intelligence Principles. (2022, January 10). International Research Center for AI Ethics and Governance. https://ai-ethics-and-governance.institute/beijing-artificial-intelligence-principles/

De Silva, D., & Alahakoon, D. (2022). An artificial intelligence life cycle: From conception to production. Patterns, 3(6), 100489. https://doi.org/10.1016/j.patter.2022.100489

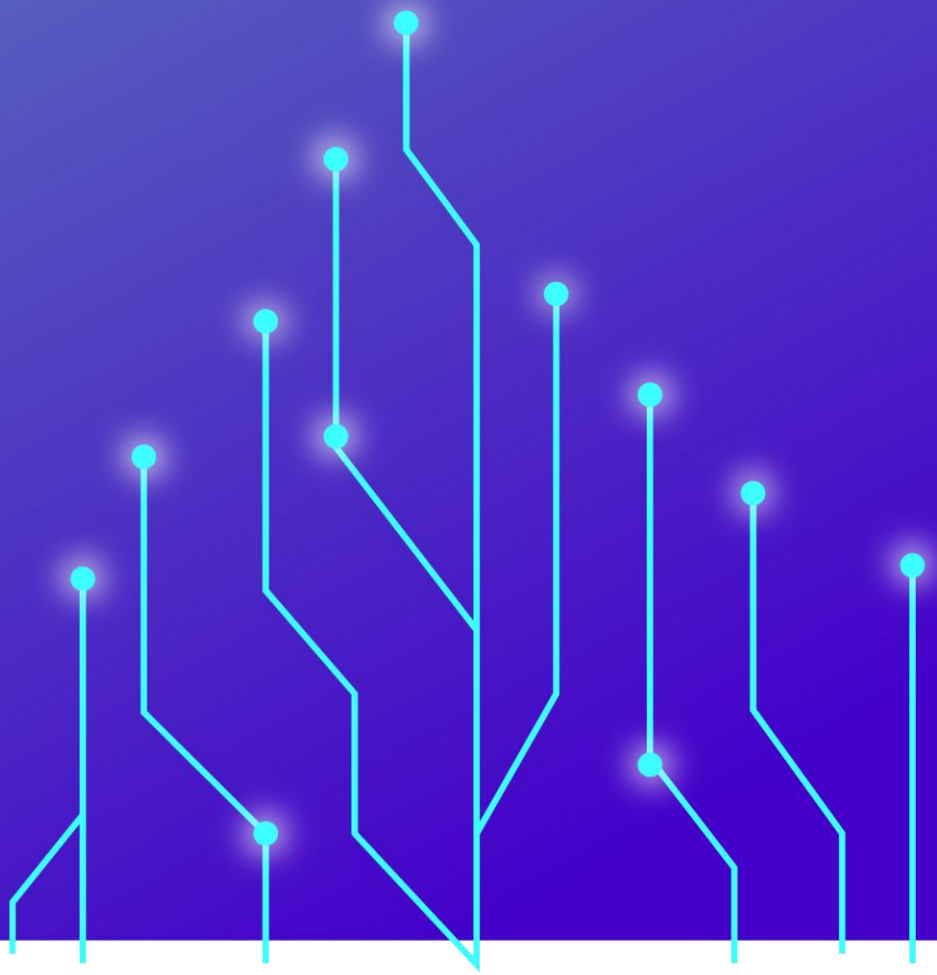
European Commission. (2023, September 12). Commission welcomes political agreement on AI Act [Text]. European Commission - European Commission. https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6473

European Data Protection Supervisor. (2024, April 9). Data Protection Officer (DPO) | European Data Protection Supervisor.

https://www.edps.europa.eu/data-protection/data-protection/reference-library/data-protection-officer-dpo_en
European Parliament. (2023, June 8). EU AI Act: First regulation on artificial intelligence. Topics | European Parliament. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review. https://doi.org/10.1162/99608f92.8cd550d1
Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5
Future of Life Institute. (2024). The AI Act Explorer | EU Artificial Intelligence Act. https://artificialintelligenceact.eu/ai-act-explorer/
High-Level Expert Group on AI (AI HLEG). (2019). Ethical Guidelines of Trustworthy AI. High-Level Expert Group on Artificial Intelligence. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
High-Level Expert Group on AI (AI HLEG). (2020). Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. European Commission.
Hillard, A., & Gulley, A. (2023, December 9). What is the EU AI Act? https://www.holisticai.com/blog/eu-ai-act
Hotz, N. (2018, September 10). What is CRISP DM? Data Science Process Alliance. https://www.datascience-pm.com/crisp-dm-2/
Immersant Data Solutions. (2023, June 9). DevOps: Bridging the Gap between Development and Operations | LinkedIn. https://www.linkedin.com/pulse/devops-bridging-gap-between-development-operations/
Interaction Design Foundation - IxDF. (2021, October 27). What is Co-Creation? Interaction Design Foundation—IxDF. The Interaction Design Foundation. https://www.interaction-design.org/literature/topics/co-creation
Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2
Junklewitz, H., Hamon, R., André, A., Evas, T., Soler Garrido, J., & Sanchez Martin, J. I. (2023). Cybersecurity of artificial intelligence in the AI Act: Guiding principles to address the cybersecurity requirement for high risk AI systems. Publications Office of the European Union. https://data.europa.eu/doi/10.2760/271009
Leijnen, S., Aldewereld, H., van Belkom, R., Bijvank, R., & Ossewaarde, R. (2020). An agile framework for trustworthy AI. 75–78.
Miller, G. J. (2022). Stakeholder roles in artificial intelligence projects. Project Leadership and Society, 3, 100068. https://doi.org/10.1016/j.plas.2022.100068

Moore, S. (2023, December 18). What is...a Corporate Legal and Compliance team? | LinkedIn. https://www.linkedin.com/pulse/what-isa-corporate-legal-compliance-team-steven-moore-iczqf/
Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. Science and Engineering Ethics, 26(4), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5
OECD. (2019). AI-Principles Overview. OECD AI Principles Overview. https://oecd.ai/en/principles
Peter, A. (2024, October 2). AI Software Architecture: Navigating the Essentials of Business App Development | LinkedIn. https://www.linkedin.com/pulse/ai-software-architecture-navigating-essentials-business-alexx-peter-uugnc/
Prem, E. (2023). From ethical AI frameworks to tools: A review of approaches. AI and Ethics, 3(3), 699–716. https://doi.org/10.1007/s43681-023-00258-9
Rochel, J., & Evéquoz, F. (2021). Getting into the engine room: A blueprint to investigate the shadowy steps of AI ethics. AI & SOCIETY, 36(2), 609–622. https://doi.org/10.1007/s00146-020-01069-w
Rodrigues, R. (2020). Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. Journal of Responsible Technology, 4, 100005. https://doi.org/10.1016/j.jrt.2020.100005
Saltz, J. (2023, June 1). What is the AI Life Cycle? Data Science Process Alliance. https://www.datascience-pm.com/ai-lifecycle/
Sanin, A. (2020, September 4). Co-Creation How-To's. Design Globant. https://medium.com/design-globant/co-creation-how-tos-e25696f56d6f
UNESCO. (2023a). Ethical impact assessment. A tool of the Recommendation on the Ethics of Artificial Intelligence. UNESCO. https://doi.org/10.54678/YTSA7796
UNESCO. (2023b). Readiness assessment methodology. A tool of the Recommendation on the Ethics of Artificial Intelligence. UNESCO. https://doi.org/10.54678/YHAA4429
UNESCO. (2023c, April 21). Artificial Intelligence: Examples of ethical dilemmas | UNESCO. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics/cases
White, C. (2023, November 30). What Does a UX Designer Actually Do? [2024 Guide]. https://careerfoundry.com/en/blog/ux-design/what-does-a-ux-designer-actually-do/
Zhou, J., & Chen, F. (2022). AI ethics: From principles to practice. AI & SOCIETY. https://doi.org/10.1007/s00146-022-01602-z
Zhou, J., Chen, F., Berry, A., Reed, M., Zhang, S., & Savage, S. (2020). A Survey on Ethical Principles of AI and Implementations. 2020 IEEE Symposium Series on Computational Intelligence (SSCI), 3010–3017. https://doi.org/10.1109/SSCI47803.2020.9308437
Zwikael, O., & Meredith, J. R. (2018). Who's who in the project zoo? The ten core project roles. International Journal of Operations & Production Management, 38(2), 474–492. https://doi.org/10.1108/IJOPM-05-2017-0274

# CU2 | Privatlivets fred og bekvemmelighed i AI



# Indhold

Comentado [KM5]: Must be updated in the end.

## 1. Introduktion

I vores digitale tidsalder har den hurtige udvikling af kunstig intelligens (AI) forandret mange ting i vores dagligdag og ændret, hvordan vi interagerer med teknologien og med hinanden. Helt centralt i denne forandring er det dynamiske samspil mellem to vigtige faktorer: beskyttelse af privatlivets fred og bekvemmelighed. I denne enhed ser vi nærmere på det komplekse forhold mellem disse to faktorer og undersøger de udfordringer og muligheder, de giver i vores smart teknologi-tidsalder.

I takt med at vi i stigende grad integrerer kunstig intelligens i vores personlige og professionelle liv, bliver det nødvendigt at se kritisk på balancen mellem at forbedre brugervenligheden og beskytte privatlivets fred. I denne kompetenceenhed (CU) får deltagerne derfor en omfattende forståelse af, hvordan privatlivets fred og bekvemmelighed sameksisterer harmonisk inden for AI-applikationer. Der lægges vægt på at fremhæve både applikationernes potentiale til at forbedre hverdagen og de etiske overvejelser, de måtte medføre.

Deltagerne vil få indsigt i grundlæggende principper for databeskyttelse, de etiske dilemmaer, som AI skaber, og de juridiske rammer, der former fremtidens teknologianvendelse. Ved at undersøge den reelle anvendelse af AI og nogle hypotetiske scenarier sigter dette kursus mod at udstyre eleverne med værktøjer til at navigere ansvarligt i AI-landskabet, som fortsat udvikler sig, og sikre, at teknologiske fremskridt forbedrer menneskets situation uden at dets rettigheder krænkes.

Denne rejse gennem AI, privatlivets fred og bekvemmelighed vil udfordre deltagerne til at tænke kritisk over teknologiens rolle i samfundet og deres ansvar som fremtidige ledere inden for AI.

BILLEDKILDE | Genereret af DALL-E

## TIPS OG ANBEFALINGER TIL UNDERVISERE

**Et engagerende spørgsmål eller overraskende fakta for at fange elevernes interesse med henblik på at forklare, hvordan AI-landskabet ser ud i dag, og med henblik på at understrege betydningen af privatlivets fred og bekvemmelighed.**

### Overraskende fakta

En nylig undersøgelse viste, at AI-systemer kan træffe beslutninger om dine jobansøgninger, kreditvurderinger og endda din sundhed – ofte uden dit direkte input eller din viden. Denne integration af AI rejser vigtige spørgsmål: Hvor meget kontrol har vi over vores data, og hvad betyder det for vores privatliv? Lad os dykke ned i AI's komplekse verden, hvor bekvemmelighed kan få forrang på bekostning af vores privatliv.

### Engagerende spørgsmål

Vidste du, at en gennemsnitlig person interagerer med AI mere end 20 gange dagligt uden at være klar over det? AI kan være fuldstændig ubemærket integreret i vores liv – i alt lige fra personlige shoppinganbefalinger til smart home-enheder, der styrer vores daglige rutiner. Men hvor ofte overvejer du, hvilke personlige oplysninger du må afgive, for at opnå disse bekvemmeligheder? Lad os undersøge, hvad der virkelig er på spil.



**Indholdet kan også præsenteres som en afstemning.**

### Afstemningsspørgsmål

"Hvilke tjenester bruger du, som du ved indsamler dine data for at forbedre funktionaliteten og personliggøre din oplevelse?"

### Valgmuligheder

- Smartphones
- Sociale medieplatforme
- Streamingtjenester
- E-handel-websites
- Smart home-enheder
- Fitness-trackere og sundhedsapps
- Stemmeassistenter
- Navigations- og rejseapps

- Bank- og finansapps
- E-mail-services

**Opfølgende spørgsmål**

"Hvor tryg er du ved, at disse tjenester indsamler dine data?"

Meget tryg || Mindre tryg || Neutral || Lidt utryg || Meget utryg

Denne tilgang hjælper med at klarlægge deltagernes bevidsthed om dataindsamling på tværs af forskellige services og sammenfatte deres følelser om privatlivets fred og brug af personlige data, hvilket giver et omfattende overblik over deltagernes holdninger til databeskyttelse.

## 2. Sarahs dilemma: kompromiset mellem privatlivets fred og bekvemmelighed

BILLEDKILDE | Genereret af DALL-E

Forestil dig Sarah, en universitetsstuderende, der lige er flyttet ind i sin første lejlighed. Sarah er ivrig efter at strømline sit travle liv og investerer i flere smarte enheder: en smart højttaler til at styre sin tidsplan og afspille musik, en smart termostat til at styre lejlighedens temperatur og et smartwatch til at overvåge hendes fitnessrutiner.

## TIPS OG ANBEFALINGER TIL UNDERVISERE

**Tip: Stil spørgsmål på afgørende tidspunkter for at gøre situationen relaterbar.**


BILLEDKILDE | Genereret af DALL-E



En morgen foreslår Sarahs smart højttaler en ny rute til hendes universitet, hvilket sparer hende tid i en periode med tæt trafik – et perfekt eksempel på bekvemmelighed. Senere får hun en personlig reklame på sin telefon for en kaffebar langs den nye rute, som frister hende med hendes yndlingslatte. Sarah bliver først glad, men undrer sig,

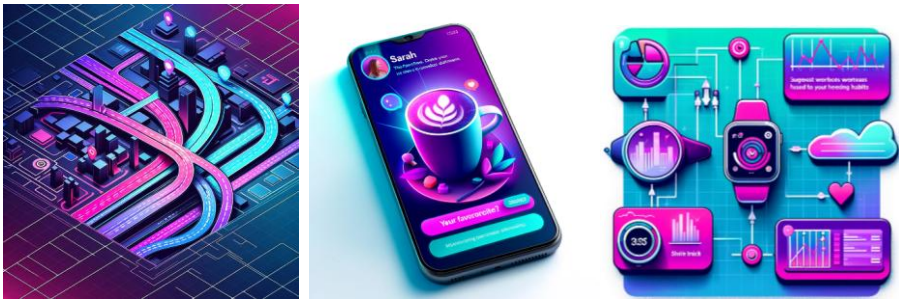
**"Hvordan vidste den det?"**

**Spørgsmål**

"Tror du, at Sarah overvejede konsekvenserne for privatlivets fred, før hun købte disse enheder? Hvorfor eller hvorfor ikke?"

**Spørgsmål**

- Hvordan har du det med, at enheder træffer beslutninger baseret på dine daglige rutiner? Hvad er de potentielle kompromiser i forhold til beskyttelse af privatlivets fred?
- Hvilke oplysninger kan være blevet delt for at personliggøre denne reklame? Er du tryg ved dette niveau af datadeling til personaliserede reklamer?



BILLEDKILDE | Genereret af DALL-E

Som dagene går, observerer Sarah flere af disse hændelser. Hendes smartwatch foreslår træningsplaner baseret på hendes fitnessniveau og nylige køb af fastfood, som er logget af hendes betalingsapp. Hendes termostat tilpasser sig hendes tidsplan og ser ud til at kunne forudse usædvanlige ændringer i hendes rutiner, som f.eks. at blive hjemme, når hun er syg.

**Spørgsmål**

"Hvornår bliver personalisering *for* indgribende? Hvor ville du selv trække grænsen?"



BILLEDKILDE | Genereret af DALL-E

Mens Sarah sætter pris på den bekvemmelighed, som disse AI-drevne enheder tilbyder, og som gør hendes liv mere problemfrit og effektivt, begynder hun at føle sig utryg ved, hvor mange personlige oplysninger de indsamler, og hvordan de bruges. Hun har ikke udtrykkeligt sagt ja til at dele sin placering eller sin

købshistorik. Sarahs glæde over bekvemmelighederne bliver hurtigt præget af bekymring for hendes privatliv.

Denne historie introducerer os til den skrøbelige balance mellem privatlivets fred og bekvemmelighed i AI. Efterhånden som AI-teknologier bliver mere og mere integreret i vores hverdag, tilbyder de en hidtil uset bekvemmelighed, men giver også anledning til væsentlige bekymringer om privatlivets fred. Udfordringen ligger i at finde en balance, hvor Sarah og brugere som hende kan nyde godt af fordelene ved AI uden at føle, at deres personlige oplysninger kompromitteres.

BILLEDKILDE | Genereret af DALL-E

**Spørgsmål:**

- Har teknologien nogensinde fået dig til at føle det samme som Sarah? Hvad gjorde du, hvis du altså gjorde noget?
- Hvad ville du råde Sarah til at gøre i denne situation?
- Hvilke skridt kan hun tage for at få bedre styr på sine data?



## 3. Forståelse af privatlivets fred i AI

Privatlivets fred kan forstås bredt som den enkeltes ret til at holde sine personlige oplysninger ude af offentlighedens søgelys og at kunne have kontrol over sine data og interaktioner. Denne ret er anerkendt som en grundlæggende menneskerettighed i mange retssystemer og internationale traktater, der understreger vigtigheden af personlig autonomi og værdighed.

Inden for AI bliver privatlivets fred endnu mere kompleks. AI-systemer er ofte afhængige af store mængder data for at lære og for at kunne træffe beslutninger. Disse data kan omfatte følsomme personlige oplysninger, som, hvis de håndteres forkert, kan føre til betydelige brud på privatlivets fred. Integration af AI i hverdagsteknologier – fra personlige assistenter og kropsbåren teknologi (*wearables*) til forudsigende analyser i sundhedsvæsenet – betyder, at beskyttelse af privatlivets fred kræver både traditionelle databeskyttelsesforanstaltninger og nye tilgange til at tackle de unikke udfordringer, som AI-teknologier udgør.

*"Privatlivets fred er en grundlæggende menneskerettighed, der er anerkendt i FN's menneskerettighedserklæring, den internationale konvention om borgerlige*

*og politiske rettigheder og mange andre globale og regionale traktater. I forbindelse med AI omfatter privatlivets fred mange vigtige dimensioner."*

Betydningen af privatlivets fred i forbindelse med AI har mange facetter:

- **Tillid**: Effektiv beskyttelse af privatlivets fred er afgørende for at bevare brugernes tillid til AI-teknologier. Uden tillid kan brugerne være tilbageholdende med at tage nye ting i brug, som kunne være nyttige for dem.
- **Etiske forpligtelser**: Der er en etisk forpligtelse til at respektere og beskytte den enkeltes privatliv. Det indebærer at sikre, at AI-systemer ikke krænker personlige rettigheder og er designet med privatlivets fred for øje fra starten.
- **Overholdelse af regler**: Mange jurisdiktioner har strenge databeskyttelsesregler, såsom GDPR i Europa, som pålægger organisationer, der behandler persondata gennem AI-systemer, specifikke forpligtelser.

## 3.1. Databeskyttelse: Beskyttelse af persondata mod uautoriseret adgang



Databeskyttelse inden for AI handler om at beskytte persondata mod uautoriseret adgang, brug eller videregivelse. AI-systemer skal være designet til at sikre, at data og følsomme oplysninger håndteres sikkert og kun tilgås af autoriserede parter. Effektive databeskyttelsesforanstaltninger hjælper med at forhindre identitetstyveri, økonomisk svindel og risici for udnyttelse af persondata. De omfatter:

- **Kryptering af data**: Kryptering af data, både når de blot oplagres, og når de bruges for at forhindre uautoriseret adgang.
- **Adgangskontrol**: Begrænsning af dataadgang til kun de personer, der har brug for den til at udføre deres arbejde.
- **Anonymisering af data**: Fjernelse af personligt identificerbare oplysninger fra datasæt, der bruges i AI, for at reducere risikoen for brud på privatlivets fred.

For eksempel kan en online detailvirksomhed indsamle kundedata til ordrebehandling og personlig markedsføring. Virksomheden bruger kryptering til at beskytte kundernes betalingsoplysninger og personlige oplysninger og sikrer dermed beskyttelse af kundernes data. Derudover implementerer de strenge interne politikker, der begrænser medarbejdernes adgang til disse følsomme data til kun at omfatte dem, der har brug for dataene til at behandle ordrer. De sikrer også kundernes samtykke gennem klare opt-in-procedurer, før de udsender markedsføringsmateriale.

## 3.2. Informationssikkerhed: Foranstaltninger til beskyttelse af dataintegritet og -fortrolighed

Informationssikkerhed er tæt forbundet med databeskyttelse, men fokuserer mere på de beskyttelsesforanstaltninger, der sikrer, at data forbliver nøjagtige, troværdige og kun tilgængelige for autoriserede brugere. I forbindelse med AI er det:

**Integritet**: Beskyttelse af data mod uautoriserede ændringer, der kan kompromittere dataenes nøjagtighed.

**Fortrolighed**: Sikring af, at personlige oplysninger holdes hemmelige og kun er tilgængelige for personer, der er autoriseret til at få adgang til dem.

**Sikkerhedsprotokoller**: Implementering af robuste protokoller som *secure socket layers* (SSL), firewalls og systemer til registrering af uautoriseret indtrængen for dermed at beskytte dataene mod eksterne angreb.

For eksempel bruger en udbyder af sundhedsydelser et elektronisk patientjournalsystem til at gemme patientdata. For at beskytte disse data anvender udbyderen flere lag af sikkerhedsforanstaltninger. Det omfatter stærk kryptering af datalagring og -overførsel, multifaktorgodkendelse (MFA) for at få systemadgang og regelmæssige sikkerhedsrevisioner for at identificere og afbøde eventuelle huller. Denne praksis er med til at sikre integriteten og fortroligheden af patientdata dermed imødegå cybertrusler.



## 3.3. Personlig autonomi: Retten til at kontrollere information om sig selv

Personlig autonomi i den digitale tidsalder, især inden for AI, handler om at bevare kontrollen over personlige oplysninger og de beslutninger, der træffes på baggrund af disse oplysninger. AI-systemer skal designes til at respektere og styrke den personlige autonomi ved:

- **Informeret samtykke**: At sikre, at brugerne er fuldt informeret om, hvordan deres data vil blive brugt, og indhentning af deres samtykke, før der indsamles data.
- **Gennemsigtighed**: At give brugerne klare oplysninger om databehandlingspraksis og logikken bag AI-beslutninger.
- **Kontrolmekanismer**: At give brugerne mulighed for at kontrollere, hvordan deres oplysninger bruges, og mulighed for at fravælge, at deres data bruges – præcis når de ønsker det.

For eksempel kan en fitness-tracking-app indsamle brugernes fysiske aktivitetsdata for at give personlige fitness-råd. Appen respekterer menneskets personlige autonomi ved tydeligt at informere brugerne om, hvilke typer data den indsamler, og hvordan dataene vil blive brugt. Den giver også brugerne mulighed for at få adgang til deres data, rette eventuelle unøjagtigheder og fravælge bestemte app-funktioners indsamling af data. Denne tilgang sikrer, at brugerne har kontrol over deres personlige oplysninger og de beslutninger, der træffes.

## 4. Databeskyttelse

Databeskyttelse handler om at beskytte persondata mod uautoriseret adgang, brug eller offentliggørelse. Det omfatter sikring af, at enkeltpersoner har kontrol over deres oplysninger, og at de bruges i overensstemmelse med deres præferencer og lovbestemmelser.

### 4.1. Vigtigheden af databeskyttelse i AI



Databeskyttelse er altafgørende inden for AI, hvor der behandles store mængder persondata. Uden tilstrækkelig beskyttelse kan enkeltpersoners følsomme oplysninger være sårbare over for brud og misbrug, hvilket kan føre til f.eks. identitetstyveri, økonomisk tab eller diskrimination. Desuden er AI-systemer stærkt afhængige af data for at kunne træffe informerede beslutninger, hvilket betyder, at datakvaliteten og -integriteten har direkte indflydelse på disse systemers ydeevne og troværdighed.

### 4.2. Rollen af forordninger som GDPR

Forordninger som f.eks. EU's generelle forordning om databeskyttelse (GDPR) er afgørende for at håndhæve principperne om databeskyttelse i AI. GDPR stiller strenge krav til indsamling, behandling og opbevaring af persondata, herunder indhentning af udtrykkeligt samtykke fra enkeltpersoner, implementering af passende sikkerhedsforanstaltninger og tilvejebringelse af mekanismer, så de registrerede kan få adgang til og kontrollere deres data. Ved at håndhæve disse regler har GDPR til formål at beskytte enkeltpersoners ret til privatliv og holde organisationer ansvarlige for håndtering af persondata.

Som eksempel – et sundhedsapp-scenarie: Forestil dig en sundhedsapp, der indsamler og gemmer brugernes medicinske data, f.eks. deres sygehistorie, diagnoser og behandlingsplaner. Denne app udgør en betydelig risiko for

brugernes databeskyttelse uden korrekt samtykke eller uden nogen form for sikkerhedsforanstaltninger. Hvis appen f.eks. mangler robust kryptering eller adgangskontrol, kan den være sårbar over for databrud og gøre følsomme medicinske oplysninger tilgængelige for uautoriserede parter. Hvis appen desuden deler brugerdata med tredjeparter uden udtrykkeligt samtykke eller anonymisering, kan den overtræde regler om beskyttelse af personlige oplysninger som f.eks. GDPR, hvilket kan få juridiske konsekvenser. Dette scenarie illustrerer vigtigheden af at implementere strenge databeskyttelsesforanstaltninger i AI-applikationer for at beskytte enkeltpersoners følsomme oplysninger og sikre overholdelse af lovkrav.

## 4.3. Vigtige aspekter af databeskyttelse

Databeskyttelse er afgørende for en moderne digital praksis, især med udbredelsen af datadrevne teknologier og services. Her er nogle centrale elementer i databeskyttelse, som hjælper med at beskytte personlige oplysninger og sikre en etisk håndtering af data:

- **Samtykke:** Samtykke indebærer, at man indhenter tilladelse fra enkeltpersoner, før man indsamler, bruger eller deler deres data. Samtykket skal gives frivilligt, være specifikt, informeret og utvetydigt. Det betyder, at enkeltpersoner skal være helt klar over, hvad de giver samtykke til, og klart forstå omfanget og konsekvenserne af databehandlingsaktiviteterne.
- **Begrænsning af formål:** Dette princip dikterer, at data skal indsamles til specifikke, eksplicitte og legitime formål og ikke viderebehandles på en måde, som er modstridende. Formålsbegrænsning hjælper med at sikre, at data kun bruges på måder, som den registrerede forventer og giver sit samtykke til, og forhindrer, at data bruges på en ondsindet eller uansvarlig måde.
- **Minimering af data:** Dataminimering betyder, at kun de data, der er nødvendige for behandlingen, indsamles og behandles. Dette princip opfordrer organisationer til at begrænse dataindsamlingen til det, der er direkte relevant og nødvendigt for at opnå et bestemt formål. Det reducerer risikoen for skader som følge af databrud, da færre data kan blive kompromitteret.
- **Ansvarlighed:** Ansvarlighed henviser til organisationers forpligtelse til at tage ansvar for de data, de behandler, og vise, at de overholder alle principper for databeskyttelse. Dette omfatter registrering af databehandlingsaktiviteter, implementering af databeskyttelsesforanstaltninger og visning af, hvordan overholdelse opnås. Dette princip sikrer, at organisationer er gennemsigtige omkring deres datahåndteringspraksis og kan holdes ansvarlige for deres databehandlingsaktiviteter.

- **Gennemsigtighed:** Gennemsigtighed kræver, at al information og kommunikation i forbindelse med behandling af personoplysninger er let tilgængelig og forståelig for enkeltpersoner. Privatlivspolitikker og meddelelser om dataindsamling skal være skrevet i et klart og tydeligt sprog. Gennemsigtighed skaber tillid, øger retfærdigheden og giver brugerne mulighed for at træffe informerede beslutninger om deres data.
- **Dataadgang og rettelser:** Personer har ret til at få adgang til deres data og til at få ukorrekte data om dem rettet. At give personer mulighed for at få adgang til og opdatere deres data efter behov sikrer datanøjagtighed og giver enkeltpersoner mulighed for at bevare kontrollen over deres oplysninger, hvilket er grundlæggende for personlig autonomi.
- **Sikkerhed:** Sikkerhed indebærer implementering af passende tekniske og organisatoriske foranstaltninger for at beskytte persondata mod utilsigtet eller ulovlig ødelæggelse, tab, ændring, uautoriseret offentliggørelse eller adgang til dataene. Dette omfatter praksisser som kryptering, sikker datalagring og kontrolleret fysisk og digital adgang.

Tilsammen udgør disse aspekter af databeskyttelse en robust ramme, som organisationer kan bruge til at beskytte persondata og dermed respektere enkeltpersoners ret til privatlivets fred. De hjælper med at opbygge en kultur, hvor der er bevidsthed om privatlivets fred, og som er i samklang med juridiske standarder og etiske forventninger i den digitale tidsalder.



## 5. Forståelse af bekvemmelighed i AI

I forbindelse med AI henviser bekvemmelighed til den lethed og effektivitet, hvormed opgaver kan udføres eller oplevelser forbedres ved hjælp af AI-teknologier. AI strømliner processer, mindsker en eventuel manuel indsats og giver personlige oplevelser, der er skræddersyet til individuelle præferencer.

Bekvemmelighed, som er muliggjort af AI-teknologier, omfatter den gnidningsløse integration af automatisering, personalisering og effektivitet på tværs af forskellige domæner. Fra sundhedsvæsenet og finansverdenen til vores ganske almindelige hverdag optimerer AI-drevne løsninger visse processer, styrker beslutningstagning og beriger vores oplevelser, hvilket i sidste ende øger produktiviteten og forbedrer menneskers generelle ve og vel.

Inden for AI er bekvemmelighed den fuldstændige integration af teknologi for at kunne strømline opgaver og øge effektiviteten. AI-teknologier udnytter algoritmer og *machine learning* til at automatisere processer, reducere manuelt arbejde og skræddersy oplevelser til folks individuelle præferencer. AI-systemer optimerer arbejdsgange ved at analysere store mængder data og lære af mønstre, hvilket sparer tid og ressourcer og samtidig giver bedre resultater. Nedenfor er et par punkter, hvor AI bidrager til bekvemmelighed på flere områder:

- **Automatisering af rutineopgaver:** AI udmærker sig ved at kunne automatisere gentagne og hverdagsagtige opgaver, som plejer at kræve menneskelig indgriben. For eksempel kan AI-drevet software håndtere planlægning af aftaler, besvarelse af kundeforespørgsler eller automatisk håndtering af e-mails. Det fremskynder processen og frigør menneskelige ressourcer til mere komplekse og kreative opgaver og øger dermed produktiviteten og effektiviteten.
- **Personliggørelse:** AI-systemer kan analysere store mængder data for at skræddersy services og produkter til individuelle præferencer og behov. I forbrugersammenhænge kan dette komme til udtryk ved, at AI anbefaler produkter baseret på tidligere køb eller seervaner, som det ses i online detailhandel og streamingtjenester. På et mere personligt plan kan AI justere varmesystemer i hjemmet baseret på husejerens tidsplan og præferencer, hvilket sikrer komfort og optimerer energiforbruget.
- **Forbedret beslutningstagning:** AI forbedrer beslutningstagning ved at give enkeltpersoner og organisationer overblik over analyser af store datasæt, som ville være for komplekse for mennesker at behandle manuelt. I sektorer som f.eks. sundhedsvæsenet hjælper AI-algoritmer med at diagnosticere sygdomme ved at analysere røntgenbilleder hurtigere og ofte mere præcist end mennesker. I finansverdenen kan AI-systemer med stor nøjagtighed og hastighed analysere markedsdata for at give indsigt i investeringer eller opdage svigagtige transaktioner.
- **Effektivitet og ressourcestyring:** AI forbedrer den samlede effektivitet betydeligt og optimerer ressourceforvaltningen. For eksempel kan AI-drevne logistik- og supply chain management-systemer hurtigt forudsige efterspørgslen, optimere leveringsruter og styre lagerbeholdningen, hvilket reducerer spild og omkostninger. På samme måde kan intelligente el-net, der drives af AI, styre elforsyningen i en by mere effektivt og tilpasse sig ændringer i efterspørgslen i realtid.
- **Tilgængelighed og brugervenlighed:** AI-teknologier gør ofte teknologien mere tilgængelig og lettere at bruge for en bredere vifte af mennesker, herunder handicappede. AI-baserede stemmeassistent-enheder hjælper personer med nedsat syn eller fysiske begrænsninger med at interagere med teknologien og styre deres omgivelser mere effektivt. AI kan også nedbryde sprogbarrierer via oversættelsestjenester i realtid, hvilket gør information og kommunikation tilgængelig for personer, der ikke har sproget som modersmål.

## 5.1. Eksempler på bekvemmelighed i det virkelige liv

**Sundhedsvæsenet**: Inden for sundhedsvæsenet gør AI-teknologier det nemmere at automatisere diagnostiske processer, analysere medicinske billeder og opdage mønstre i patientdata for at hjælpe sundhedspersonalet med at diagnosticere sygdomme hurtigere og mere præcist. Det sparer sundhedspersonale og patienter for tid, hvilket fører til mere effektive sundhedsydelser og bedre resultater for patienterne.

BILLEDKILDE | Genereret af DALL-E

**Bekvemmelighed i sundhedsvæsenet via AI**

- **Automatiserede diagnostiske processer**: AI nedsætter tidsforbruget ved medicinsk dataanalyse, som f.eks. billedoptagelse og diagnostiske tests, hvilket giver mulighed for hurtigere diagnoser.
- **Forbedret diagnostisk nøjagtighed**: AI-algoritmer opdager subtile mønstre i data, som mennesker måske overser, hvilket øger diagnosernes nøjagtighed og troværdighed.
- **Forudsigende analyser**: AI bruger data fra forskellige kilder, herunder kropsbåren teknologi, til at forudsige potentielle sundhedsproblemer, før de bliver alvorlige, hvilket fremmer forebyggende sundhedspleje.
- **Strømlinet patientadministration**: AI optimerer hospitalets arbejdsgange ved at styre tidsplaner, patientindtag og dirigering til de rigtige behandlingssteder, hvilket reducerer ventetiden og forbedrer patientoplevelsen.
- **Personlige behandlingsplaner**: Algoritmer analyserer individuelle patientdata for at skræddersy behandlingsplaner, der med stor sandsynlighed vil være effektive, baseret på patientens unikke sundhedsprofil.



**Finansverdenen:** AI-drevne finansielle værktøjer giver bekvemmelighed ved at tilbyde personlig investeringsrådgivning baseret på individuelle finansielle mål, risikoprofil og markedstendenser. Disse værktøjer bruger algoritmer til at analysere store mængder finansielle data og give anbefalinger til investeringsstrategier, hvilket hjælper enkeltpersoner med at træffe informerede beslutninger og optimere deres finansielle porteføljer.

BILLEDKILDE | Genereret af DALL-E

**Bekvemmelighed i finansverdenen via AI**

- **Personlig investeringsrådgivning**: AI-værktøjer analyserer den enkeltes finansielle mål, risikotolerance og personlige økonomiske situation for at give skræddersyede investeringsanbefalinger.
- **Avanceret dataanalyse**: Disse systemer bruger algoritmer til at gennemgå store mængder finansielle data, herunder markedstendenser og historiske resultater, for at identificere de bedste investeringsmuligheder.
- **Optimeret porteføljestyring**: AI hjælper med dynamisk at justere investeringsporteføljer baseret på markedsændringer i realtid og individuelle økonomiske mål, hvilket forbedrer det potentielle afkast, samtidig med at man har kontrol over risikoeksponeringen.
- **Effektiv markedsindsigt**: AI-værktøjer giver hurtig indsigt i markedsforholdene og hjælper investorer med at forstå komplekse markedsdynamikker og træffe rettidige beslutninger.
- **Risikovurdering**: Algoritmer evaluerer den risiko, der er forbundet med forskellige investeringsmuligheder, så brugerne kan træffe informerede valg baseret på deres risikotolerance.
- **Automatisering af transaktioner**: AI kan automatisere handels- og investeringstransaktioner, strømline udførelsesprocessen og reducere sandsynligheden for menneskelige fejl.
- **Forbedret sikkerhed**: Avancerede sikkerhedsforanstaltninger, som f.eks. detektering af uregelmæssigheder, anvendes af AI til at identificere og afbøde potentielle trusler mod investeringskonti.

**I hverdagen:** Smart home-enheder, som er drevet af AI, såsom stemmeassistenter, smarte termostater og automatiserede sikkerhedssystemer til hjemmet, giver bekvemmelighed ved at automatisere husholdningsarbejdet og forbedre komfort og sikkerhed. Disse enheder lærer brugernes præferencer at kende over tid og justerer indstillingerne i overensstemmelse hermed, hvilket giver en perfekt og personlig oplevelse, der forenkler de daglige rutiner og forbedrer folks livskvalitet.

BILLEDKILDE | Genereret af DALL-E

**Bekvemmelighed i hverdagen via AI**

- **Automatisering af husholdningsopgaver**: AI-drevne enheder såsom robotstøvsugere og automatiserede køkkenmaskiner kan håndtere rutineopgaver og frigør dermed tid for brugerne.
- **Personlige indstillinger**: Enheder såsom intelligente termostater og belysningssystemer lærer og tilpasser sig brugerens præferencer og justerer automatisk indstillingerne for at opnå den optimale komfort.
- **Stemmeassisteret teknologi**: Stemmeassistenter som Alexa og Google Assistant giver håndfri kontrol over husholdningsapparater og gør det lettere at få adgang til information og styre opgaveløsningen.
- **Forbedret sikkerhed i hjemmet**: Intelligente sikkerhedssystemer bruger AI til at overvåge hjemmet, genkende kendte ansigter, opdage usædvanlige aktiviteter og advare husejere om potentielle trusler.
- **Energieffektivitet**: AI-drevne enheder optimerer energiforbruget i hjemmet ved at lære spidsbelastningstider og justere driften for at reducere spild, hvilket sænker el-regningen.
- **Smidig integration**: Smart home-systemer generelt forbinder forskellige enheder, så de kan arbejde smidigt sammen og forbedre brugervenligheden gennem integreret kontrol og interaktion.
- **Fjernovervågning og -styring**: Boligejere kan overvåge og fjernstyre smart home-enheder via smartphones, hvilket giver ro i sindet, når de er væk hjemmefra.

Selvom AI øger bekvemmeligheden betydeligt, giver det også anledning til vigtige overvejelser, såsom bekymringer om privatlivets fred, behovet for menneskelig overvågning af systemet og den potentielle fare for jobfortrængning i specifikke sektorer (dvs. at AI-drevne enheder overtager menneskers job). At sikre, at AI-systemer bruges etisk og ansvarligt, er afgørende for at maksimere deres fordele og samtidig minimere potentielle skadelige

virkninger. Konklusionen er, at bekvemmelighed i AI handler om at udnytte potentialet i kunstig intelligens til at gøre livet lettere, sjovere og mere effektivt. Efterhånden som AI-teknologier udvikler sig og integreres i forskellige aspekter af livet, vil deres potentiale til at skabe større bekvemmelighed sandsynligvis udvides, hvilket indvarsler hidtil usete effekter på tværs af alle samfundets facetter.

## TIPS OG ANBEFALINGER TIL UNDERVISERE

Præsenter hypotetiske scenarier, der udfordrer deltagerne til at vælge mellem at prioritere privatlivets fred eller bekvemmelighed.

**Eksempel**

**Hypotetisk scenarie: Sundhedstracker-app**

**Baggrund**: Du overvejer at få en ny sundhedstracker-app, HealthMate. Denne app tilbyder personlige fitness- og kostplaner ved at analysere dine daglige aktiviteter, dit madindtag og dine sundhedsmålinger.

**Scenarieudfordring:** HealthMate kan synkronisere med dine sociale medier for at indsamle flere livsstilsdata for at opnå en forbedret personalisering og dele information med andre partnere, der kan sende dig målrettede annoncer baseret på dine sundhedsmål.

**Spørgsmål til refleksion:**

- Ville du vælge den forbedrede personalisering, hvis du vidste, at dine data kunne blive brugt til målrettet reklame?
- Hvordan afvejer du fordelene ved en skræddersyet sundhedsplan i forhold til risikoen ved at dele dine data?

Dette scenarie får deltagerne til at vurdere balancen mellem bekvemmeligheden ved skræddersyede sundhedsanbefalinger og risiciene for krænkelse af privatlivets fred ved datadeling.



## 6. Samspillet mellem privatlivets fred og bekvemmelighed

Det dynamiske samspil mellem privatlivets fred og bekvemmelighed i moderne teknologi giver både muligheder og udfordringer. I takt med at teknologien i stigende grad integreres i vores hverdag, bliver det afgørende at forstå, hvordan man afbalancerer disse aspekter. Nedenfor beskrives konsekvenserne af afvejninger, etiske dilemmaer og konsekvenser i den virkelige verden af at afbalancere privatlivets fred med bekvemmelighed.

**Afvejninger mellem privatlivets fred og bekvemmelighed:**

At afbalancere privatlivets fred og bekvemmelighed kræver omhyggelig navigering mellem de iboende kompromiser:

- Begrænsninger i dataindsamling: Implementering af beskyttelse af privatlivets fred betyder ofte en begrænsning af mængden og typen af data, der indsamles og bruges. Det kan begrænse funktionaliteten i AI-systemer, der er afhængige af store datasæt for at forbedre serviceydelsen og brugeroplevelsen.
- Afgivelse af privatlivets fred til fordel for bekvemmelighed: På den anden side indebærer øget bekvemmelighed ofte indsamling og analyse af omfattende mængder af personlige data. Det kan være sporing af brugernes adfærd, præferencer og hvor de befinder sig, hvilket kan krænke privatlivets fred.

**Etiske dilemmaer, der opstår som følge af disse afvejninger:**

Balancen mellem privatlivets fred og bekvemmelighed bringer flere etiske dilemmaer frem i lyset:

- Kompromis med privatlivets fred: Hvor meget privatliv er den enkelte villig til at give afkald på til fordel for øget bekvemmelighed? Svaret varierer ofte ud fra individuelle opfattelser og den kontekst, teknologien bruges i.
- Praksis for dataindsamling: Organisationer er også etisk ansvarlige for at indsamle, bruge og dele personlige oplysninger. Etiske overvejelser skal styre beslutninger om, hvilke data der indsamles, hvordan de bruges, og hvor meget brugerne informeres om denne praksis.



**Konsekvenser i den virkelige verden af at vælge bekvemmelighed frem for privatliv eller omvendt:**

De beslutninger, der træffes om balancen mellem privatliv og bekvemmelighed, har reelle effekter:

- Konsekvenser ved at prioritere bekvemmelighed: At vælge bekvemmelighed øger ofte risici som databrud, identitetstyveri og betydelige krænkelser af privatlivets fred. Det kan føre til langvarig skade på brugernes tillid, og det kan have juridiske konsekvenser for virksomhederne.
- Konsekvenser ved at prioritere privatlivets fred: Omvendt kan en høj prioritering af privatlivets fred begrænse effektiviteten af specifikke teknologiske løsninger. Det kan føre til mindre personaliserede tjenester og potentielt bremse den teknologiske innovation og udvikling, der kan være til gavn for samfundet.

At navigere i krydsfeltet mellem privatlivets fred og bekvemmelighed kræver en nuanceret tilgang, der tager højde for teknologiens etiske, sociale og personlige konsekvenser. Organisationer og enkeltpersoner skal stræbe efter at finde en mellemvej, der respekterer brugernes privatliv, samtidig med at de udnytter de

fordele, teknologien kan give. Denne balance er ikke statisk og skal løbende evalueres, efterhånden som teknologien udvikler sig, og samfundets værdier skifter. Forståelse af denne dynamik hjælper interessenter med at træffe informerede beslutninger, der beskytter individuelle rettigheder og samtidig omfavner den digitale tidsalders fremskridt.

## 7. Vigtigheden af privatlivets fred og bekvemmelighed

Privatlivets fred og bekvemmelighed i AI er afgørende for brugernes tillid og anvendelse af teknologi. Efterhånden som AI integreres i vores hverdag, er det afgørende at opnå en balance mellem beskyttelse af personlige data og forbedring af brugeroplevelsen. Privatlivets fred sikrer beskyttelse af personlige oplysninger, mens bekvemmelighed forbedrer effektivitet og personalisering, når man anvender teknologierne. Udfordringen ligger i at harmonisere dette forhold for at fremme en etisk udbredelse af AI og samfundsmæssig accept og sikre, at fremskridt inden for AI giver fordele uden at gå individuelle rettigheder kompromitteres.

**Forøgelse af brugertillid og teknologiadoption**

Prioritering af privatlivets fred og bekvemmelighed i AI-teknologier er afgørende for at opbygge brugernes tillid til systemerne og fremme en bredere anvendelse af dem. Når AI-systemer håndterer privatlivets fred på en gennemsigtig måde og giver problemfri og praktiske brugeroplevelser, bliver de lettere accepteret af brugerne. Her kommer nogle eksempler på, hvordan prioritering af disse aspekter kan påvirke brugernes tillid til og anvendelse af teknologien:

- **Opbygning af tillid**: Brugere har en tendens til at stole på teknologi, der respekterer deres privatliv og forbedrer deres dagligdag uden væsentlig indblanding. AI-løsninger, der effektivt beskytter brugerdata og samtidig tilbyder personlige og brugbare services, har en tendens til at opnå et positivt omdømme blandt brugerne.
- **Tilskyndelse til ibrugtagning**: Når brugere oplever, at en AI-teknologi forbedrer deres produktivitet eller bekvemmelighed uden at gå på kompromis med deres privatliv, er de mere tilbøjelige til at tage den i brug og anbefale den til andre. Når værdien af en AI-teknologi viser sig gennem øget bekvemmelighed, kombineret med stærk beskyttelse af privatlivets fred, motiveres brugerne til at integrere disse teknologier i deres dagligdag.

**Juridiske og etiske overvejelser ved implementering af AI**

Implementering af AI-teknologier kræver omhyggelig overvejelse af både juridiske og etiske konsekvenser for at sikre en ansvarlig brug af dem og overholdelse af diverse regler:

- **Overholdelse af lovgivningen**: Love som den generelle forordning om databeskyttelse (GDPR) indeholder strenge retningslinjer for databeskyttelse, som AI-teknologier, der opererer i eller er rettet mod brugere i EU, skal følge. Overholdelse af disse regler handler ikke kun om, at det er en juridisk nødvendighed, men også om at vise, at der tages højde for og hensyn til brugernes privatliv.

- **Etiske rammer**: Ud over juridiske krav styrer forskellige etiske rammer udviklingen af AI for at sikre, at teknologierne bruges til gavn for samfundet. Disse rammer hjælper udviklere og virksomheder med at navigere i komplekse spørgsmål som bias, retfærdighed, gennemsigtighed og ansvarlighed i AI-systemer.

**Påvirkning af samfundsnormer og den enkeltes adfærd**

AI's rolle i at balancere mellem privatlivets fred og bekvemmelighed har stor indflydelse på samfundets normer og den enkeltes adfærd:

- **Skabelse af forventninger til og præferencer for privatlivets fred**: Efterhånden som AI-teknologier bliver mere integrerede i hverdagen, påvirker de, hvordan enkeltpersoner opfatter og værdsætter deres privatliv. Det kan føre til ændrede forventninger, hvor nogle brugere bliver mere bevidste om opretholdelse af privatlivets fred, mens andre kan blive mere ligeglade med hensyn til at dele personlige oplysninger.
- **Indflydelse på daglige rutiner og beslutningstagning**: AI-teknologier, der tilbyder bekvemmelighed, kan ændre dagligdagen markant. Smart home-enheder, personaliserede indkøbsoplevelser og automatisering af éns personlige økonomistyring kan ændre hverdagens aktiviteter, gøre livet lettere og skabe afhængighed af teknologien i forbindelse med beslutningstagning.



Ved at forstå og håndtere det komplekse samspil mellem privatlivets fred og bekvemmelighed kan AI-udviklere og -brugere øge accepten af teknologien og sikre, at den bidrager positivt til samfundsudviklingen og den enkeltes ve og vel. Disse overvejelser er afgørende for en bæredygtig udvikling af AI-teknologier i et samfund, der i stigende grad er bevidst om privatlivets fred og etiske standarder.

## 8. At navigere mellem risici og fordele

Potentielle risici forbundet med AI-teknologier i forhold til privatlivets fred og bekvemmelighed omfatter:

- **Brud på datasikkerheden:** AI-systemer håndterer ofte store mængder følsomme data, hvilket gør dem til mål for ondsindede angreb. Databrud kan resultere i uautoriseret adgang til personlige oplysninger, hvilket fører til identitetstyveri, økonomisk tab og skade på ens omdømme.
- **Tab af privatlivets fred:** AI-teknologier kan indsamle og analysere persondata uden den enkeltes samtykke eller viden, hvilket kompromitterer retten til privatlivets fred. Det kan underminere tilliden mellem brugere og AI-systemer og føre til problemer med datamisbrug og -udnyttelse.

- **Bias i AI-algoritmer:** AI-algoritmer kan fastholde bias i deres træningsdata, hvilket fører til diskriminerende resultater. Biased AI-algoritmer kan resultere i uretfærdig behandling eller beslutningstagning, hvilket kan gøre samfundsmæssige uligheder større og underminere tilliden til AI-systemer.

**Konsekvenser af disse risici:**

Disse risici kan underminere tilliden til AI-systemer og have negative konsekvenser for enkeltpersoner og samfundet:

- **Underminering af tillid:** Brud på datasikkerheden og krænkelser af privatlivets fred kan underminere tilliden mellem brugere og AI-systemer, hvilket fører til mindre udbredelse og brug af systemerne. Enkeltpersoner kan blive tilbageholdende med at dele deres data eller engagere sig i AI-teknologier, hvilket hindrer udnyttelse af teknologiernes potentielle fordele.
- **Negative samfundsmæssige konsekvenser:** Biased AI-algoritmer kan fastholde diskrimination og ulighed, hvilket fører til uretfærdig behandling af individer og social uro. Det kan skade den sociale sammenhængskraft og tilliden til institutioner og instanser samt forværre splittelser i samfundet.
- **Juridiske og lovgivningsmæssige konsekvenser:** Brud på datasikkerheden og krænkelser af privatlivets fred kan få juridiske og lovgivningsmæssige konsekvenser for organisationer, der er ansvarlige for AI-systemer. Bøder, retssager og skader på omdømmet kan have betydelige økonomiske og driftsmæssige konsekvenser.



**Potentielle fordele:**

På trods af risiciene giver AI-teknologier mange potentielle fordele:

- **Øget effektivitet:** AI strømliner processer og automatiserer opgaver, hvilket øger effektiviteten og produktiviteten. Ved at håndtere ensformige opgaver frigør AI tid, så medarbejdere kan fokusere på andre og vigtigere aktiviteter.
- **Bedre datadrevne beslutninger:** AI-algoritmer analyserer store mængder data for at afdække viden og mønstre, så organisationer kan træffe bedre informerede beslutninger. AI-dreven viden kan optimere driften, forbedre kundeoplevelser og fremme innovation.
- **Forbedrede brugeroplevelser:** AI-personalisering forbedrer brugeroplevelsen ved at levere skræddersyede anbefalinger og services. AI forbedrer brugernes tilfredshed og engagement i mange forhold, lige fra personaliserede produktanbefalinger til forudsigende vedligeholdelse i et produktionsanlæg.

**Afhjælpning af risici:**

Organisationer bør implementere robuste sikkerhedsforanstaltninger for at mindske disse risici, sikre gennemsigtighed og ansvarlighed i AI-systemer og prioritere etiske overvejelser i hele AI-udviklingens livscyklus. Ved proaktivt at håndtere disse risici kan organisationer opbygge tillid til AI-systemer og minimere potentiel skade på enkeltpersoner og samfundet.

For at navigere effektivt i krydsfeltet mellem risici og fordele ved AI-teknologier kan organisationer anvende følgende strategier:

- **Implementere robuste sikkerhedsforanstaltninger:** Organisationer bør implementere robuste sikkerhedsforanstaltninger for at beskytte mod brud på datasikkerheden og uautoriseret adgang til data. Det omfatter kryptering, adgangskontrol og regelmæssige sikkerhedsrevisioner.
- **Sørge for gennemsigtighed og ansvarlighed:** Gennemsigtighed og ansvarlighed er afgørende for at opbygge tillid til AI-systemer. Organisationer bør være åbne om, hvordan AI-teknologier bruges, og sikre, at organisationen agerer ansvarligt.
- **Prioritere etiske overvejelser:** Etiske overvejelser bør være styrende for udvikling og implementering af AI. Organisationer bør prioritere retfærdighed, gennemsigtighed og ansvarlighed i AI-systemer for at mindske bias og sikre etisk brug af data.

Ved at indføre en proaktiv risikostyring og ansvarlig praksis for implementering af AI kan organisationer maksimere fordelene ved AI-teknologier og samtidig minimere potentielle skadelige virkninger på enkeltpersoner og samfundet.



## 9. Casestudie: Overvågning ved hjælp af AI

### Scenarie

Lokale myndigheder implementerer AI-drevne overvågningssystemer i en travl by for at forbedre den offentlige sikkerhed. Disse systemer bruger avancerede teknologier som ansigtsgenkendelse, adfærdsanalyse og objektdetektering til at overvåge det offentlige rum, identificere potentielle trusler og reagere på farlige situationer i realtid.

BILLEDKILDE | Genereret af DALL-E

**Sikkerhedsforbedringer vs. brud på privatlivets fred:**

| Forbedringer af sikkerheden | Brud på privatlivets fred |
| --- | --- |
| AI-overvågningssystemerne styrker sikkerhedstiltag ved at gøre det muligt for de retshåndhævende myndigheder hurtigt at opdage og reagere på mistænkelige aktiviteter eller potentielle trusler mod sikkerheden.<br><br>Teknologi til ansigtsgenkendelse hjælper med at identificere personer på overvågningslister eller personer, der er meldt savnet, hvilket gør det lettere at gribe ind og værne om offentlighedens sikkerhed. | På trods af sikkerhedsfordelene opstår der bekymring om de potentielle brud på privatlivets fred, der er forbundet med AI-overvågning. Kontinuerlig overvågning og dataindsamling i det offentlige rum rejser spørgsmål om enkeltpersoners ret til privatliv og frihed under konstant overvågning.<br><br>Der er risiko for fejlidentifikation og falske positiver, hvilket kan føre til uretmæssige anklager og krænkelser af privatlivets fred for folk, der egentlig er uskyldige, men som systemet har udpeget som mistænkelige |



**Vigtigheden af en gennemtænkt AI-implementering:**

Eksemplet med overvågning ved hjælp af AI understreger vigtigheden af en gennemtænkt AI-implementering for at afbalancere sikkerhedsbehov med hensynet til privatlivets fred:

- **Klare politikker og regler:** En gennemtænkt AI-implementering indebærer, at der etableres klare politikker og regler for overvågningsteknologier for at sikre gennemsigtighed, ansvarlighed og overholdelse af love om privatlivets fred.

- **Etisk brug af data:** Organisationer skal prioritere etisk brug af data, som er indsamlet gennem overvågningssystemer, respektere individers ret til privatliv og minimere risikoen for brud på privatlivets fred eller misbrug af personlige oplysninger.
- **Algoritmisk gennemsigtighed og ansvarlighed:** Implementering af gennemsigtige AI-algoritmer og mekanismer for opretholdelse af ansvarlighed sikrer, at overvågningssystemerne er retfærdige, nøjagtige og ansvarlige for deres handlinger, hvilket mindsker risikoen for bias eller fejl, der kan føre til krænkelser af privatlivets fred.
- **Offentligt engagement og tilsyn:** At inddrage offentligheden i diskussioner om AI-overvågning og involvere interessenter i beslutningsprocessen fremmer gennemsigtighed, tillid og ansvarlighed og fremmer ansvarlig implementering og brug af overvågningsteknologier.

Afslutningsvis understreger eksemplet om overvågning ved hjælp af AI behovet for en gennemtænkt implementering af AI, hvor der er balance mellem sikkerhedsforbedringer og hensynet til privatlivets fred. Organisationer kan implementere overvågningssystemer, der forbedrer sikkerheden, samtidig med at de prioriterer etiske principper, gennemsigtighed og ansvarlighed, respekterer individers ret til privatlivets fred og fremmer offentlighedens tillid til systemerne.

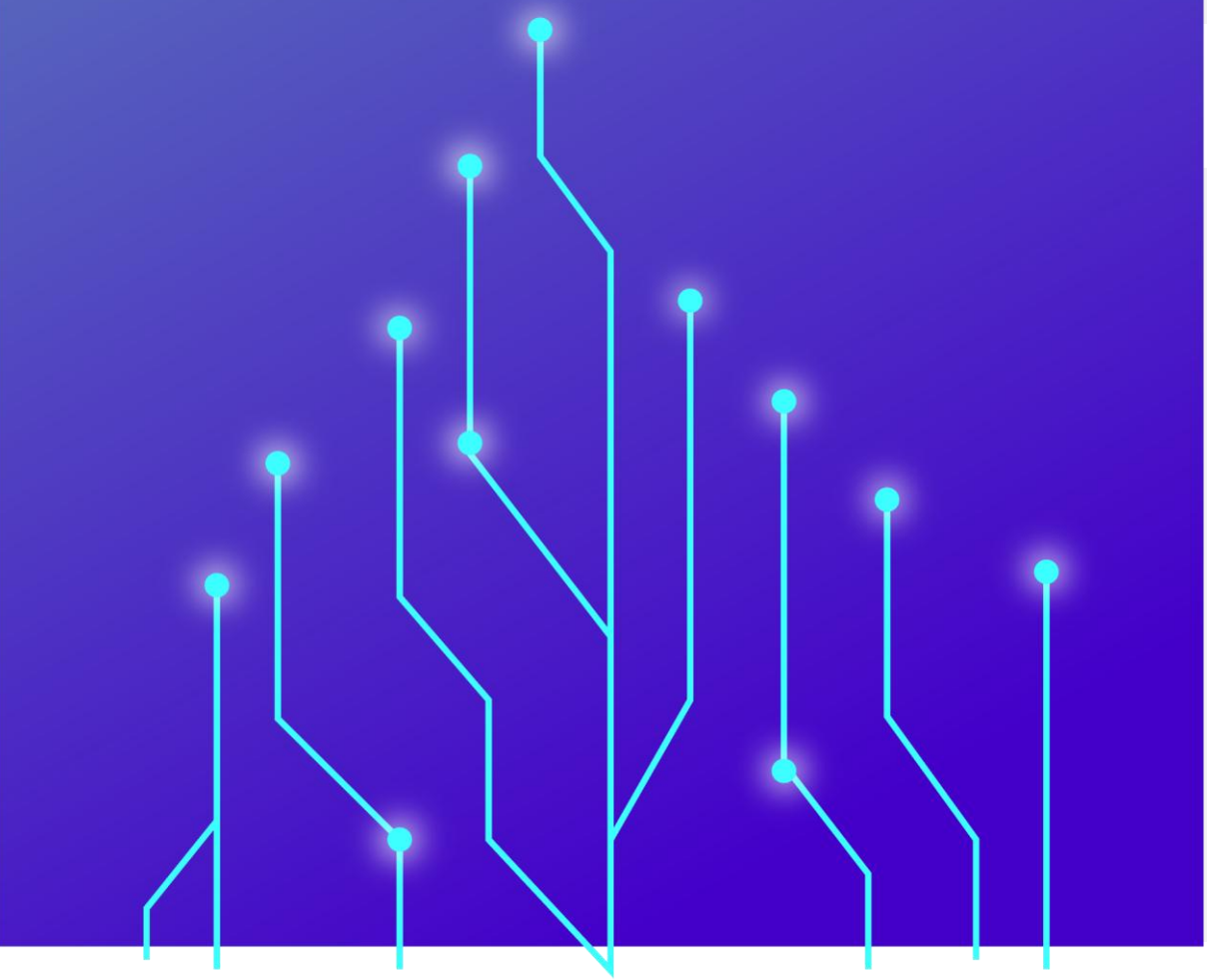
## 10. Konklusion



At se nærmere på privatlivets fred og bekvemmelighed inden for kunstig intelligens (AI) understreger den vigtige og nødvendige balance med hensyn til at fremme den generelle tillid til og bredere anvendelse af teknologierne. Efterhånden som AI gennemsyrer mange forskellige aspekter af sundhedsvæsenet, finansverdenen og dagliglivet, giver det betydelige fordele som f.eks. øget effektivitet og personaliserede services, men dette ledsages af et behov for robust beskyttelse af privatlivets fred og grundige, etiske overvejelser. Overholdelse af juridiske rammer som GDPR og håndtering af etiske dilemmaer er afgørende for at sikre en ansvarlig anvendelse af AI.

Efterhånden som AI er med til at forme samfundsnormer og individuel, menneskelig adfærd, bliver det afgørende at opretholde en balance mellem brugervenlighed og databeskyttelse. Udviklere, brugere og politiske beslutningstagere skal samarbejde for at sikre, at AI-teknologier forbedrer livet og hverdagen uden at gå på kompromis med grundlæggende rettigheder. At lægge vægt på gennemsigtighed og ansvarlighed er nøglen til at udnytte AI's potentiale på en ansvarlig måde og sikre, at det forbliver en gavnlig og troværdig teknologi i vores stadigt mere digitale verden.

# CU3 I Algoritmer og deres begrænsninger



# Indhold

Comentado [KM6]: Must be updated after end of proof reading session.

## 1. Introduktion[1]

Denne håndbog er designet til at være et centralt værktøj for undervisere, der afholder Charlie Project-kurset. Den beskriver ikke kun kursets grundlæggende indhold, men giver også detaljerede instruktioner, tips og anbefalinger til underviserne om, hvordan de effektivt kan gennemføre kurset. Derudover vil håndbogen foreslå en række aktiviteter, der har til formål at engagere deltagerne og forbedre deres læringsoplevelse. Den indeholder også retningslinjer for, hvordan man evaluerer deltagernes læringsresultater og sikrer, at kursets mål opfyldes, og at både undervisere og deltagere har en klar forståelse af de forventede kompetencer, der skal udvikles i løbet af kurset. Denne omfattende tilgang vil hjælpe underviserne med at skabe et dynamisk og effektivt læringsmiljø.[2]

I denne enhed introduceres de grundlæggende koncepter for algoritmer, deres modeller og begrænsninger. Deltagerne vil få en forståelse af den rolle, algoritmer spiller i forskellige sektorer, og de potentielle faldgruber, der opstår ved deres brug. Emnerne omfatter algoritmisk kompleksitet, optimering og begrænsningerne i algoritmisk beslutningstagning.

I dagens digitale tidsalder spiller algoritmer en stadig større rolle i udformningen af forskellige aspekter af vores liv, lige fra det indhold, vi bruger online, til de beslutninger, der træffes af automatiserede systemer. Det er afgørende for alle, der skal navigere i dette landskab, at forstå den indviklede algoritmiske kompleksitet, optimeringsstrategier og de begrænsninger, der ligger i algoritmiske beslutningsprocesser.

At forstå **algoritmisk kompleksitet** indebærer, at man undersøger algoritmers effektivitet og performance, når de løser forskellige opgaver. Det indebærer at forstå de ressourcer - såsom tid og hukommelse - som algoritmer kræver for at løse problemer, og hvordan disse krav skaleres i takt med større inputstørrelser. Denne forståelse giver os mulighed for at vurdere levedygtigheden og



[1] Bemærk: Under udarbejdelsen af dette dokument anvendte forfatterne Large Language Models (LLM) til at hjælpe med korrekturlæsning af teksten, strukturering af indholdet og identifikation af relevante eksempler i materialet. Efter at have brugt dette AI-værktøj har forfatterne omhyggeligt gennemgået og forfinet indholdet efter behov. Forfatterne påtager sig det fulde ansvar for indholdets integritet og nøjagtighed i den endelige publikation.

[2] Til underviseren: Denne håndbog har til formål at guide dig gennem materialet til kursusenheden. De understøttende PowerPoint-filer indeholder oplysninger, der svarer til dem, som e-læringen dækker, men i mere detaljeret form. PowerPoint-filerne er omfattende og er ikke beregnet til at blive gennemgået i deres helhed under de interaktive sessioner. I stedet kan du spørge i begyndelsen af de interaktive sessioner, hvilke emner der var sværest at forstå for de studerende, og kun bruge disse supportslides under sessionen. Du kan spørge om dette ved at bruge værktøjer som Mentimeter eller Kahoot og tilføje kursusemnerne fra kompetenceenhedens indhold, der er beskrevet i punktform i slutningen af introduktionsafsnittet. Desuden er det meningen, at de kompetence-enheds-baserede casestudier også skal indgå i den interaktive session. Alle yderligere forslag i håndbogen er kun vejledende; du er velkommen til at bruge dem, som du finder det passende.

anvendeligheden af at bruge specifikke algoritmer i virkelige sammenhænge (Cormen et al., 2009).

**Optimeringsmetoder** udgør et andet grundlæggende aspekt af forståelsen af algoritmer. Disse metoder har til formål at forfine algoritmernes effektivitet ved at justere deres parametre eller omstrukturere deres grundlæggende logik. Uanset om det drejer sig om at reducere beregningstiden, maksimere ressourceudnyttelsen eller optimere til bestemte formål, kan anvendelse af passende optimeringsstrategier forbedre algoritmernes ydeevne betydeligt på tværs af forskellige domæner (Cormen et al., 2009).

Men ud over de potentielle fordele ved algoritmisk beslutningstagning er der iboende begrænsninger, som kræver nøje overvejelse. Algoritmer fungerer inden for foruddefinerede rammer og er afhængige af kvaliteten af inputdata og nøjagtigheden af de underliggende antagelser. Bias i data, fejlbehæftede antagelser i modelleringerne eller helt usædvanlige tilfælde (*edge cases*) kan føre til fejlagtige resultater eller have utilsigtede konsekvenser, hvilket fremhæver skrøbeligheden ved algoritmiske beslutningssystemer.

At få en større forståelse af algoritmisk kompleksitet, optimeringsstrategier og de begrænsninger, der ligger i algoritmiske beslutningsprocesser, kan fremme en dybere forståelse af algoritmiske spidsfindigheder og udfordringer. Det giver enkeltpersoner evnen til at tænke kritisk, hvilket er nødvendigt for at vurdere algoritmiske løsninger, identificere potentielle faldgruber og advokere for en ansvarlig og etisk algoritmisk praksis i en stadig mere algoritmedrevet verden.



## TIPS OG ANBEFALINGER TIL UNDERVISERE

### Tips og anbefalinger til undervisere om at arbejde med de grundlæggende algoritmebegreber

**1. Afklar centrale begreber**
Før du bevæger dig ud i diskussioner, skal du sikre dig, at alle deltagere har en klar forståelse af grundlæggende terminologi som f.eks. "algoritmisk kompleksitet", "optimering" og "algoritmers begrænsninger". Det kan ske gennem en kort præsentation eller en ordliste i begyndelsen af sessionen.

**2. Forhold dig til applikationer i den virkelige verden**
Forbind abstrakte begreber med deres reelle anvendelse i forskellige sektorer såsom finansverdenen, sundhedssektoren eller e-handel. Det hjælper deltagerne med at se relevansen og konsekvenserne af algoritmer i hverdagsscenarier, hvilket gør indholdet mere engagerende og forståeligt.

**3. Brug visuelle hjælpemidler**
Algoritmer kan være komplekse og abstrakte. Brug diagrammer, flowcharts og animationer til at visualisere begreber som algoritmisk flow, kompleksitet og

optimeringsprocesser. Det kan hjælpe deltagerne med lettere at forstå og huske informationen.

**4. Opmuntr til spørgsmål**
Skab en atmosfære af åbenhed, hvor deltagerne føler sig trygge ved at stille spørgsmål. Det kan ske gennem regelmæssige Q&A-sessioner efter at have dækket hvert større emne, så det sikres, at deltagerne forstår materialet, før de går videre.

**5. Indarbejd casestudier**
Diskuter virkelige cases, hvor algoritmisk beslutningstagning er lykkedes eller er stødt på problemer. En analyse af disse cases kan sætte gang i en diskussion om de etiske implikationer, effektiviteten og de praktiske overvejelser i forbindelse med brug af algoritmer.

**6. Tilskynd til kritisk tænkning**
Tilskynd deltagerne til at tænke kritisk over begrænsningerne og de potentielle bias i algoritmiske modeller. Det kan ske gennem gruppediskussioner eller debatter om, hvordan man kan afbøde disse problemer i algoritmedesign og algoritmeimplementering.

## Anbefalede opvarmningsaktiviteter

**1. Algoritme-rollespil**
En rollespilsaktivitet, hvor deltagerne simulerer en algoritme i aktion, enten ved manuelt at sortere genstande, følge nogle simple instruktioner for at nå et mål eller ved at udspille de trin, en algoritme kan tage i beslutningsprocesser. Denne øvelse hjælper med at illustrere begrebet 'trinvis behandling i algoritmer'.



**2. Icebreaker om algoritmisk kompleksitet**
Start med en simpel opgave, som f.eks. at sortere bøger efter farve, og skaler det op til at sortere efter genre og derefter efter forfatter inden for hver genre. Diskuter, hvordan kompleksiteten øges for hvert lag, der tilføjes, og drag paralleller til algoritmisk kompleksitet, og hvordan der bruges flere ressourcer, når opgaverne bliver mere komplekse.

**3. Optimeringsudfordring**
Giv små grupper en opgave, f.eks. at arrangere stole i et rum, så der er plads til så mange mennesker som muligt, men med færrest mulige rækker. Lad dem prøve i praksis, diskutere deres strategier og derefter afsløre, hvordan optimeringsstrategier kan forbedre deres løsninger.

**4. Etisk debat**
Organiser en debat om de etiske konsekvenser af algoritmer i beslutningstagning med fokus på potentielle bias og konsekvenserne af at stole for meget på automatiserede systemer. Det kan varme deltagerne op til den kompleksitet og det ansvar, der ligger i at arbejde med algoritmer.

**5. Bryd-algoritmen-spillet**
Giv deltagerne en simpel algoritme eller et sæt regler, og udfordr dem til at finde *edge cases* eller scenarier, hvor algoritmen fejler. Denne aktivitet kan fremhæve algoritmernes begrænsninger og de potentielle fejl i den algoritmiske logik.

Disse aktiviteter og strategier vil ikke kun gøre læringsprocessen engagerende, men også give deltagerne en dybere forståelse af kompleksiteten i algoritmisk beslutningstagning.

## 2. Forstå de grundlæggende koncepter for algoritmer, deres modeller og begrænsninger

### 2.1. Algoritmisk kompleksitet

Algoritmer er i dag allestedsnærværende. De guider computere til at behandle information, løse problemer og træffe vigtige beslutninger. De kan løse problemer lige fra organisering af data til at finde de bedste rejseruter eller foreslå film. I en tid, der er domineret af big data og sociale medier, er det vigtigt at vide, hvordan algoritmer fungerer, for at forstå deres begrænsninger og potentielle bias. De forbedrer ikke bare, hvordan computere udfører opgaver; de påvirker også vores hverdagsoplevelser, og derfor spiller de en afgørende rolle i moderne databehandling.

En algoritme er en sekvens af entydige instruktioner til at løse et problem, den skal være korrekt, altid give en korrekt løsning, og den skal være endelig, den skal altså afsluttes. Det er en trinvis procedure, som viser os, hvad der skal gøres for at nå et bestemt mål. Vi kan tænke på det som madlavning: Vi følger en opskrift, der fortæller os præcis, hvilke ingredienser vi skal bruge, hvor meget af hver, og hvilke skridt vi skal tage. På samme måde guider en algoritme en computer eller en person gennem en række handlinger for at opnå et bestemt resultat, hvad enten det drejer sig om at sortere en liste med tal, finde den korteste rute på et kort eller anbefale film baseret på dine præferencer.



Begrebet algoritmer har eksisteret i lang tid og går tilbage til ældgamle civilisationer. Et af de tidligste eksempler er den euklidiske algoritme, som tilskrives den gamle græske matematiker Euklid, og som blev udviklet omkring 300 f.Kr. til at finde den største fælles divisor mellem to tal.

Gennem historien har forskellige kulturer og civilisationer udviklet deres egne algoritmer til at løse matematiske problemer, navigere i geografisk terræn eller udføre opgaver effektivt. Disse algoritmer blev ofte overleveret mundtligt eller gennem skrevne tekster.

Selve udtrykket "algoritme" kommer fra navnet på den persiske matematiker Muhammad ibn Musa al-Khwarizmi, som levede under den islamiske guldalder i det 9. århundrede. Al-Khwarizmi skrev en bog med titlen "Al-Kitab al-Mukhtasar fi Hisab al-Jabr wal-Muqabala" (som kan oversættes til: Den kortfattede bog om beregning ved hjælp af tilføjelser og udligning), som introducerede systematiske metoder til løsning af matematiske ligninger. Ordet "algoritme" stammer fra den latinske version af hans navn, "Algoritmi".

Siden da har algoritmer udviklet sig og er blevet grundlæggende inden for forskellige områder, især med fremkomsten af moderne computere. I dag bruges algoritmer ikke kun i matematik, men også i datalogi, ingeniørvidenskab, biologi,

økonomi og mange andre discipliner til at løse komplekse problemer og automatisere processer.

## 2.2. Algoritmers karakteristika

Vi vil her forsøge at definere algoritmers karakteristika. Begrebet algoritme har ikke en generelt accepteret og formel definition. Knuth (1997) kom frem til en liste med fem egenskaber, der er bredt accepteret som krav til en algoritme (da man ofte bruger den engelske term på dansk, er den i visse tilfælde medtaget nedenfor):

- ***Finiteness* (begrænsethed):** "En algoritme skal altid afsluttes efter et begrænset antal trin ... et meget begrænset antal, et rimeligt antal". Begrænsethed sikrer, at algoritmer afsluttes efter et begrænset antal trin, så man undgår uendelige sløjfer, der kan bruge beregningsressourcer i det uendelige. Det svarer til at lægge et puslespil, hvor den sidste brik til sidst falder på plads og signalerer, at algoritmen er afsluttet. Denne egenskab ved begrænsethed giver forudsigelighed og kontrol over beregningsprocesser.
- ***Definiteness* (præcision):** "Hvert trin i en algoritme skal være præcist defineret; de handlinger, der skal udføres, skal være nøje og utvetydigt specificeret for hvert enkelt trin". Dette understreger behovet for utvetydige trin, der ikke behøver yderligere fortolkning. En klar algoritme skitserer tydeligt hver handling, der skal udføres på hvert trin i problemløsningen, og sikrer, at der ikke er nogen tvetydighed eller forvirring om, hvad der skal gøres. Denne egenskab er afgørende, fordi den gør det muligt for alle, uanset deres baggrund eller ekspertise, at forstå og implementere algoritmen korrekt.
- **Input:** "... datamængder, som gives til den i starten, før algoritmen begynder. Disse input tages fra specifikke sæt af objekter". Input refererer til de data eller oplysninger, som algoritmen bruger til at producere et resultat. Dette input kan være enhver form for data, f.eks. tal, tekst, billeder eller endda andre algoritmer. Input er vigtigt, fordi det udgør det råmateriale, som algoritmen behandler og håndterer for at løse et problem eller udføre en opgave.
- **Output:** "... datamængder, som har en bestemt relation til input". Output refererer til det resultat eller den løsning, som algoritmen producerer efter at have behandlet input. Dette output kan også antage forskellige former afhængigt af arten af det problem, der skal løses, f.eks. en enkelt værdi, flere værdier, en beslutning, en visuel gengivelse af noget eller enhver anden form for information, der repræsenterer det ønskede resultat.
- ***Effectiveness* (effektivitet):** "... alle de operationer, der skal udføres i algoritmen, skal være så enkle, at de i princippet kan udføres nøjagtigt og inden for et begrænset tidsrum af en mand med papir og blyant". Effektivitet understreger algoritmers praktiske anvendelighed og gennemførlighed. Ligesom en opskrift bliver svær at følge, hvis den

indeholder ukendte ingredienser eller ingredienser, der er svære at få fat i, så skal algoritmer kunne bruge tilgængelige ressourcer til at udføre deres opgaver. Ved at sikre, at algoritmerne er praktiske og funktionsdygtige, sikrer effektiviteten i algoritmen, at der kan findes en løsning eller output inden for rammerne af de tilgængelige ressourcer og den tilgængelige tid.

Som vi kan se, er ovenstående beskrivelse af en algoritmes egenskaber måske intuitivt klar, men den mangler formel stringens, da det ikke er helt klart, hvad "præcist defineret" betyder, eller "nøje og utvetydigt specificeret" betyder, eller "skal være så grundlæggende, at" osv., så anser vi den for at være tilstrækkelig til vores formål.

**Comentado [KM7]:** In the English version, there is a spelling mistake: it should say 'We **will** consider it…' (not 'We well consider it…')

## 2.3. Algoritmens struktur

Et af hovedkoncepterne bag algoritmer er abstraktion. Abstraktion i denne sammenhæng henviser til ideen om at skjule komplekse detaljer og kun fokusere på de væsentlige aspekter, der er nødvendige for at forstå og løse et problem. Det er som at zoome ud for at se det større billede uden at fare vild i finkornede detaljer. I algoritmedesign hjælper abstraktion med at forenkle problemløsningsprocessen ved at bryde den ned i mindre, håndterbare dele. Det giver os mulighed for at tackle hver del for sig uden at skulle forstå alle de indviklede detaljer på én gang.



I algoritmer indgår typisk flere instruktioner, der 'arbejder sammen', snarere end blot ét enkelt trin. Når vi skaber komplekse programmer, udfører vi en række af disse instruktioner i en bestemt rækkefølge, gentager dem eller træffer valg baseret på specifikke forhold.

At vide, hvordan algoritmer fungerer, indebærer at forstå forskellige måder, hvorpå instruktioner kan kombineres. Disse systemer eller mønstre giver en struktureret tilgang til at organisere og sekvensere instruktionerne i en algoritme. De hjælper med at nedbryde komplekse problemer til håndterbare opgaver, hvilket gør algoritmen mere forståelig og lettere at implementere. Algoritmer består typisk af enkle strukturer, der er organiseret på en hierarkisk måde. Disse strukturer styrer flowet af instruktioner i algoritmen og dikterer, i hvilken rækkefølge de skal udføres. Dette kaldes ofte algoritmens kontrolstruktur.

En kontrolstruktur styrer i bund og grund rækkefølgen af udførelsen af instruktionerne i en algoritme. Den bestemmer, hvilken vej udførelsesprocessen skal tage, baseret på de betingelser, der er defineret i strukturen. Kontrolstrukturer er et grundlæggende aspekt af enhver algoritme, da de sikrer, at instruktionerne udføres i den rigtige rækkefølge, og at det ønskede resultat opnås.

Fælles for alle kontrolstrukturer er, at de har et enkelt indgangspunkt og et enkelt udgangspunkt. Det ene indgangspunkt markerer begyndelsen af

kontrolstrukturen, hvor udførelsesprocessen kommer ind i strukturen. På den anden side markerer det enkelte exit-punkt slutningen af kontrolstrukturen, hvor udførelsesprocessen forlader strukturen og går videre til næste del af algoritmen. Denne funktion sikrer, at udførelsesprocessen følger en defineret rute inden for kontrolstrukturen, hvilket opretholder algoritmens integritet og korrekthed.

Der er tre måder at sammensætte instruktioner på:

**- Sekventiel:** Ved denne fremgangsmåde udføres instruktioner én efter én i en lineær rækkefølge. Det er meget tæt på at følge en opskrift trin for trin. Sekventielle algoritmer er udbredte i hverdagsopgaver, som f.eks. at lave en sandwich eller udregne summen af tal. De sikrer, at handlinger sker i en bestemt rækkefølge uden afstikkere eller yderligere beslutningstagning.

Et eksempel: En algoritme til at lave en kage.
1. Forvarm ovnen.
2. Smør en kageform og drys den indvendigt med mel.
3. Rør smør og sukker sammen i en skål.
4. Pisk æggene i, et ad gangen.
5. Rør vaniljeekstrakt i.
6. Bland mel, bagepulver og salt i en anden skål.
7. Tilsæt gradvist de tørre ingredienser til den våde blanding, skiftevis med mælk.
8. Hæld dejen i den smurte kageform.
9. Sæt kageformen i den forvarmede ovn.
10. Bag i 25-30 minutter, eller indtil en tandstikker, der er stukket ind i midten, kommer ren ud.
11. Tag kagen ud af ovnen, og lad den køle af i formen i 10 minutter.
12. Server kagen.



**Comentado [KM8]:** In the English version, this step appears as no. 15 - but I assume it must be step 12.

**- Betinget eller alternativ:** Disse instruktioner indfører nogle beslutningspunkter. Baseret på visse betingelser kan programmet tage forskellige ruter. Tænk på det som et valg mellem flere muligheder. Betingede udsagn (som "if", "else" og "switch") giver os mulighed for at håndtere forskellige scenarier.

Et eksempel: En algoritme, der simulerer processen med at beslutte, om man skal tage en paraply eller en jakke med, når man forlader sit hus, baseret på vejrudsigten:

1. Tjek vejrudsigten for regn.
2. Hvis (*if*) vejrudsigten forudsiger regn, så:
3. Tag en paraply med.
4. (*Else*) Hvis vejrudsigten ikke forudsiger regn, og temperaturen er under 15 grader, så:
5. Tag en jakke med.
6. Forlad huset.

Bemærk, at instruktionerne 3 og 5 har en særlig rolle i algoritmedefinitionen. Disse instruktioner udføres kun, hvis den foregående instruktion udløser et positivt svar.

**Comentado [KM9]:** In the English version, it says something about "indentation of instructions 3 and 5", but there is no indentation in the text. I suggest improving/changing the English version and then try to translate into Danish (and the other language).

- **Gentagende eller iterativ:** Gentagelse er kernen i disse algoritmer. De udfører et sæt instruktioner gentagne gange, indtil en bestemt betingelse er opfyldt. Loops (såsom "*for*" og "*while*") falder ind under denne kategori. De er afgørende for opgaver som behandling af store datasæt, simulering af begivenheder eller løsning af optimeringsproblemer.

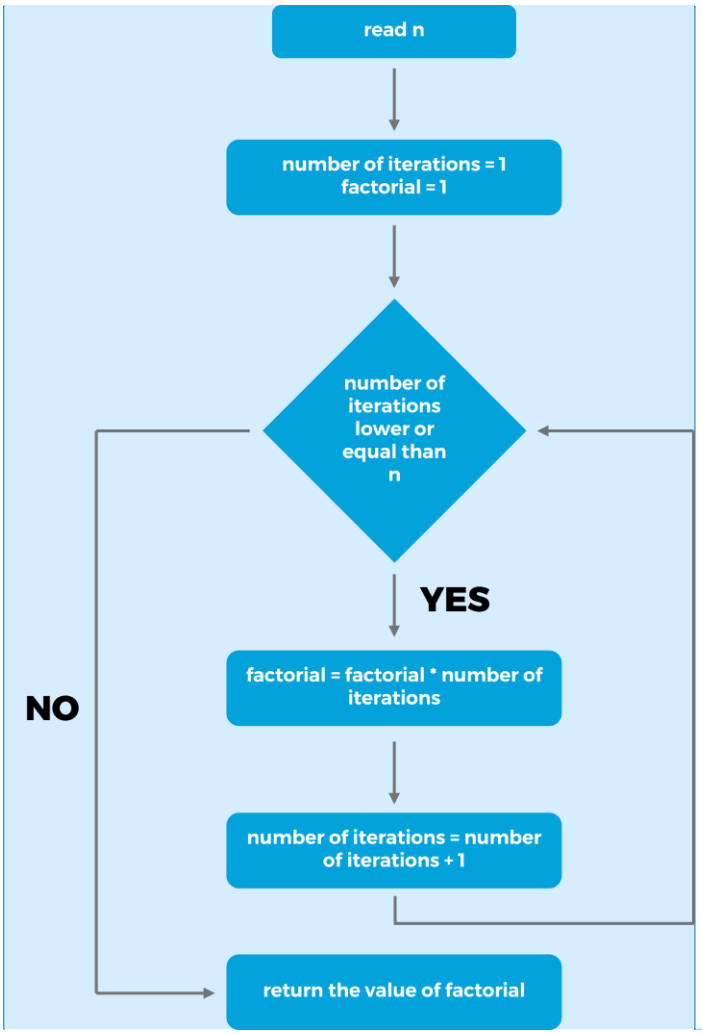
Et eksempel: Et program, der beregner fakultetet af et tal ved hjælp af en løkke. Fakultetet af et tal er en matematisk fremgangsmåde, der ganger et givet tal, n, med alle heltal mindre end n ned til 1.

1. Læs n, som er det tal, vi ønsker at beregne fakultet for.
2. Sæt værdien af fakultet til 1.
3. Loop igennem fra 1 til n:
4. Multiplicer fakultet med værdien af løkken.
5. Vis værdien af fakultetstallet.

Bemærk, at indrykningen af instruktion nummer 4 også har en specifik rolle i algoritmedefinitionen. Denne instruktion udføres n gange, én gang ved hver iteration.

**Comentado [KM10]:** I asked a programming specialist at the department because I was unsure if this was translated correctly. He said that the English version should be improved since it did not make totally sense - at least in his view. The source text needs to be better in order to make a correct translation.
Again, in the English text, it is mentioned that "the indentation of instruction number 4 also has a specific role" but there is no indentation.

**Illustration 1.** Eksempel på loop




Kilde | Egen udarbejdelse

**Comentado [KM11]:** Could student assistans help translate this? And copy it to the ppt-file?

Vi kan konstruere mere og mere komplekse algoritmer ved at kombinere disse tre teknikker. Lad os se et eksempel på en algoritme, der bruger alle tre teknikker til at finde den maksimale værdi i et heltal på en liste:

Liste med 10 heltal. Tallet 10 er det første element på listen.


1. Det aktuelle element er det første på listen.
2. Den maksimale værdi er det aktuelle element.
3. Vi kigger ikke på alle elementer i listen, men laver et loop:
4. Hvis det aktuelle element er større end den maksimale værdi, så:
5. Den maksimale værdi er det aktuelle element.
6. Det aktuelle element er det næste på listen.
7. Returnér den maksimale værdi.

**Comentado [KM12]:** I repeat my previous comment:
I asked a programming specialist at the department because I was unsure if this was machine translated correctly. He said that the English version should be improved since it did not make totally sense - at least in his view. The source text needs to be better in order to make a correct translation.

75

## 2.4. Algoritmers effektivitet

Inden for datalogi er algoritmer egentlig programmer. Et program kan defineres som beskrivelsen af en algoritme i et begrænset repertoire af maskininstruktioner. Den gruppe af programmer, som en computer har, kaldes software. For at få en computer til at udføre en opgave skal vi først udtænke og designe en algoritme til at løse den, og derefter skal vi programmere den ind i computeren. Målet er at omdanne den konceptuelle algoritme til et sæt tydelige instruktioner i et specifikt programmeringssprog, baseret på forskellige tilgange til programmeringsprocessen (programmeringsparadigmer).

Efterhånden som problemerne bliver mere komplekse, bliver de algoritmer, der skal løse dem, det også. Det er på dette tidspunkt, at det bliver nødvendigt at kunne vurdere de omkostninger, som en algoritme har. Algoritmisk effektivitet henviser til, hvor effektivt en algoritme udnytter beregningsressourcerne. Denne egenskab er afgørende for at forstå en algoritmes performance og indebærer en analyse af dens ressourceforbrug, herunder tid og plads. At opnå maksimal effektivitet indebærer at minimere ressourceforbruget. Det kan dog være kompliceret at sammenligne algoritmers effektivitet, fordi forskellige ressourcer ikke kan sammenlignes direkte.

Vi kan betragte en algoritme som effektiv, når dens ressourceforbrug, kaldet beregningsomkostninger, forbliver på eller under en acceptabel grænse. Denne grænse bestemmes ud fra algoritmens evne til at fungere inden for en rimelig tidsramme eller plads på en standardcomputer, ofte i forhold til inputstørrelsen. I bund og grund betyder "acceptabel", at algoritmen kan udføres inden for

praktiske begrænsninger, hvilket sikrer, at den forbliver levedygtig og kan bruges i den virkelige verden.

Der er to hovedmål med hensyn til algoritmisk effektivitet:

**Tidskompleksitet:** Dette er den beregningsmæssige kompleksitet, der beskriver den tid, det tager at køre en algoritme som en funktion af størrelsen på input til programmet.

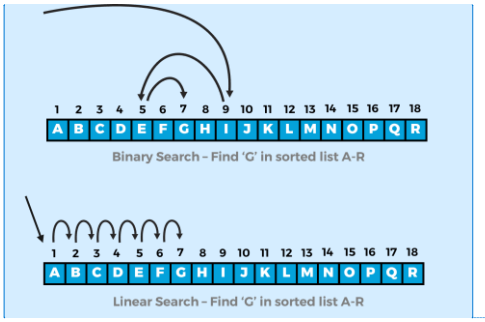
**Pladskompleksitet:** Dette er den mængde hukommelse, som en algoritme skal bruge for at køre til ende.

Målet er at minimere både tid og pladsbehov. Men der foretages ofte en afvejning mellem disse to. For eksempel kan en algoritme, der kører hurtigere, kræve mere hukommelse, og en algoritme, der bruger mindre hukommelse, kan være langsommere. Valget af algoritme afhænger af opgavens specifikke krav. Effektive algoritmer finder en balance mellem disse overvejelser og sigter mod optimal ydeevne i virkelige scenarier.

Lad os betragte et simpelt eksempel på to forskellige algoritmer, der kan bruges til at søge efter et element i en liste, og som har samme pladskompleksitet, men forskellig tidskompleksitet. Derefter vil vi analysere, hvornår man skal bruge henholdsvis den ene eller den anden.

Lineær søgning er en algoritme, der bruges til at finde en værdi i en liste, og den minder meget om algoritmen til at finde den maksimale værdi på listen, som vi udviklede før. Den ser på hvert element på listen fra begyndelsen, indtil målværdien er fundet, eller slutningen af listen er nået. Under hver iteration sammenligner algoritmen det aktuelle element med målværdien. Hvis der findes et match, afsluttes algoritmen, og målværdiens indeks returneres.

Binær søgning er en effektiv algoritme, der bruges til at finde en målværdi i en sorteret liste ved gentagne gange at dele søgeintervallet i to. Den begynder med at undersøge det midterste element på listen og sammenligner det med målværdien. Hvis det midterste element matcher målet, er søgningen vellykket. Hvis målet er mindre end det midterste element, fortsætter søgningen i den nederste halvdel af listen; hvis målet er større, flyttes søgningen til den øverste halvdel. Denne proces med at halvere søgeintervallet gentages, indtil målværdien er fundet, eller søgeintervallet er udtømt. Binær søgning fungerer efter princippet om 'del og hersk', hvilket reducerer søgeområdet betydeligt for hver iteration.




Kilde | Board Infinity

Med disse to eksempler kan vi nemt se, at der findes adskillelige algoritmer til at udføre den samme opgave. Hver algoritme har sine egne betingelser, styrker og svagheder; lineær søgning er nem at implementere og velegnet til små lister og usorterede data. Men algoritmens effektivitet falder, når listens størrelse vokser, hvilket gør den mindre anvendelig til store datasæt, for hvis det element, vi søger, er tæt ved slutningen af listen, skal vi evaluere de fleste af dens elementer. Binær søgning er meget effektiv, hvilket gør den ideel til at søge i store og sorterede datasæt. Den kræver dog, at dataene er sorteret fra starten, hvilket kan være en forudsætning for at kunne anvende den.



## 2.5. Begrænsninger

I løbet af de seneste år har vi set en bemærkelsesværdig stigning i teknologier inden for maskinlæring og kunstig intelligens, som begge er stærkt afhængige af algoritmer til at behandle information fra enorme datasæt. Denne stigning har ført til bemærkelsesværdige fremskridt inden for forskellige områder, herunder billed- og videogenkendelse og -generering, robotteknologi, medicinsk analyse, beslutningstagning, stemmekloning og sprogbehandling. Disse fremskridt illustrerer det grænseløse potentiale, som algoritmer har, og det skubber os videre ind i en æra præget af hidtil uset innovation, men det er vigtigt at forstå, at selv med disse vigtige fremskridt har algoritmer sine begrænsninger. Der er problemer, som ikke kan løses af en algoritme, og der er problemer, som vi ikke er parate til at løse.

Først vil vi beskrive de problemer, som en algoritme ikke kan løse. Der er to kategorier: Uafgørlige problemer er dem, der er teoretisk umulige at løse med nogen algoritme, og uhåndterlige problemer er dem, der ikke har nogen løsninger inden for rimelig tid.

Et klassisk eksempel på et uafgørligt problem er *halting*-problemet, altså et beslutningsproblem med et binært (ja eller nej) svar, som er uafgørligt. Det indeholder den udfordring, om et computerprogram til sidst vil stoppe eller køre i

**Comentado [KM13]:** Could student assistants translate this illustration? And copy it to the ppt-file?

det uendelige. Det kan formuleres som: "Er det muligt at skrive et computerprogram, der kan fortælle et hvilket som helst program og dets input, og uden at udføre programmet, om det nogensinde vil stoppe?

På den anden side er der de uhåndterlige problemer. Nogle uhåndterlige problemer kan løses inden for en rimelig tid og med små input, men algoritmen kan ikke skaleres til større input. Nogle uhåndterlige problemer løses hver dag, selv om de er uhåndterlige. Det gælder f.eks. ruteplanlægningsalgoritmer, tidsplanlægning og ressourceplanlægning. Disse har en tendens til at bruge en 'gættet løsning' på et uhåndterligt problem, som derefter kan afprøves relativt hurtigt. Dette kaldes heuristik, når der bruges viden og erfaring til at komme med intelligente gæt. Den heuristiske tilgang kan også forsøge at finde en passende løsning og ikke nødvendigvis den perfekte eller ideelle.

Et af de mest berømte uhåndterlige problemer er ***Travelling Salesman Problem*** (TSP). I denne udfordring er opgaven at finde den korteste rute, der gør det muligt for en sælger at besøge hver by på et kort præcis én gang, før han vender tilbage til startbyen. Målet er at minimere den samlede tilbagelagte afstand, hvilket er en stor beregningsmæssig udfordring på grund af den eksponentielle stigning i mulighederne, når antallet af byer vokser.

Lad os lave nogle tal for at se, hvordan det vokser. For at beregne antallet af veje mellem byerne skal vi beregne fakultet af antallet af byer minus én. Antallet af ruter, vi skal ud på, er halvdelen af antallet af forbindelser mellem byerne (vi skal kun bruge hver vej én gang).

**Comentado [KM14]:** Can someone check if this is the correct term in Danish? Should it be plural (fakulteterne) in Danish?

I tabellen kan vi se, hvordan antallet af muligheder vokser meget hurtigt :[3]

| Number of cities | Number of connections between cities | Number of possible routes |
|---|---|---|
| 5 | 4! = 24 | 12 |
| 10 | 9! = 362.880 | 181.440 |
| 15 | 14! = 87.178.291.200 | 43.589.145.600 |

**Comentado [KM15]:** Can student assistants translate this table? And copy it to the ppt-file?

## 2.6. Begrænsninger ved maskinlæringsalgoritmer

I æraen med maskinlæring (eller dyb læring) har vi en enorm samling af algoritmer, der er designet til at løse forskellige opgaver, f.eks: Beslutningstræer, *Random Forest*, *Support Vector Machines* eller neurale netværk (i alle dets versioner) ... De kan alle tilpasses forskellige problemer ved hjælp af en træningsproces, hvor vi viser dem data, og hvorefter de tilpasser sig. Ved hjælp af

[3] Eksempel hentet fra https://runestone.academy/ns/books/published/mobilecsp/Unit5-Algorithms-Procedural-Abstraction/Limits-of-Algorithms.html

denne proces kan vi opnå brugbare løsninger. Det er dog vigtigt at bemærke, at de også har deres begrænsninger.

- **Datakvalitet og -mængde:** Maskinlæringsalgoritmer er stærkt afhængige af kvaliteten og mængden af de data, der er tilgængelige til træningen af dem. Begrænsede eller biased data kan føre til unøjagtige eller biased resultater af modellernes svar. Derudover kan utilstrækkelige data hindre algoritmens evne til at generalisere ukendte data på en tilstrækkeligt god måde.

- **Overtilpasning og undertilpasning (*overfitting* og *underfitting*):** Maskinlæringsmodeller have en tendens til *overfitting*, hvor de lærer at opfange 'støj' eller irrelevante mønstre i træningsdataene, hvilket fører til dårligere performance, når de skal behandle nye data. På den anden side opstår *underfitting*, når modellen er for simpel, hvilket resulterer i en suboptimal performance.

- **Fortolkningsmuligheder:** Mange maskinlæringsmodeller, især komplekse modeller som neurale netværk, mangler fortolkningsmuligheder, hvilket gør det vanskeligt at forstå og stole på deres resultater. Det kan være problematisk på områder, hvor fortolkning er afgørende, f.eks. inden for sundhedssektoren og finansverdenen.

- **Beregningsmæssige ressourcer:** Træning af komplekse maskinlæringsmodeller kræver ofte betydelige beregningsressourcer, herunder meget kraftig hardware og omfattende infrastruktur til databehandling i stor skala. Begrænsede beregningsressourcer kan begrænse skalerbarheden og anvendeligheden af maskinlæringsløsninger.

- **Domænespecifik ekspertise:** Udvikling af effektive maskinlæringsløsninger kræver ofte domænespecifik ekspertise til korrekt forbehandling af data, udvikling af relevante funktioner og fortolkning af modeloutput.

- **Kontinuerlig læring og tilpasning:** Maskinlæringsalgoritmer kan have svært ved at tilpasse sig skiftende omgivelser eller en datadistribution, der hele tiden udvikler sig. Kontinuerlig overvågning og omskoling af modeller er nødvendig for at sikre, at deres performance forbliver robust og tidssvarende.



## TIPS OG ANBEFALINGER TIL UNDERVISERE

### Tips og anbefalinger til undervisere om arbejdet med algoritmekompleksitetskoncepter

**1. Byg et stærkt udgangspunkt**
Start med en grundlæggende oversigt over, hvad algoritmer er, og deres historiske udvikling for at sætte det hele ind i en kontekst. Det hjælper med at

opbygge en fortælling, der gør det abstrakte koncept mere håndgribeligt og relaterbart.

**2. Brug analogier og metaforer**
Sammenlign algoritmer med hverdagsaktiviteter såsom madlavning eller at finde vej i en by. Det kan hjælpe med at afmystificere konceptet og gøre det tilgængeligt for elever, der måske ikke har en stærk teknisk baggrund.

**3. Lav interaktive fremvisninger**
Vis algoritmer i aktion gennem live-kodningssessioner eller interaktive webværktøjer, der giver deltagerne mulighed for at indtaste eksempler og se, hvordan algoritmerne behandler dem. Dette kan være ret engagerende og lærerigt.

**4. Forsøg at fremme interaktiviteten i klasseværelset**
Opfordr deltagerne til at stille spørgsmål og inviter dem til at diskutere for at give dem mulighed for at udtrykke deres forståelse af eller forvirring om de emner, der bliver behandlet. Det hjælper også med at vurdere deres engagement og generelle forståelse i realtid.

**5. Trinvis læring**
Bryd komplekse emner ned i mindre, overskuelige dele. Start med grundlæggende begreber, og introducer gradvist mere kompleksitet. Denne trinvise tilgang hjælper med at forebygge kognitiv overbelastning.

## Opvarmningsaktiviteter

**1. Enkel oprettelse af algoritmer**
Bed deltagerne om at skrive en simpel algoritme til en hverdagsopgave, f.eks. at lave te eller gøre sig klar til at gå på arbejde. Denne aktivitet får dem til at tænke i de trinvise, logiske sekvenser, som er grundlæggende for algoritmedesign.

**2. Historisk kig på algoritmer**
Diskuter den euklidiske algoritme i gruppen, og undersøg dens enkelte trin, og hvordan den historisk set er blevet brugt til at løse forskellige udfordringer. Dette kan føre til en dybere forståelse af, hvordan grundlæggende begreber i algoritmer har dybe rødder.

**3. Rollespil vedrørende kontrolstruktur**
Lav et rollespil, hvor hver deltager spiller en del af en kontrolstruktur i en algoritme, f.eks. en betingelse eller et loop. Det vil hjælpe dem med at visualisere og forstå, hvordan forskellige dele af en algoritme interagerer.

**4. Algoritme-fejlfindingsudfordring**
Giv deltagerne en simpel algoritme med bevidste fejl, og bed dem om at "debugge" den. Det kan være en sjov og engagerende måde at uddybe deres forståelse af, hvordan algoritmer er opbygget og fungerer.

## Tips og anbefalinger til vurdering af deltagernes læring



**1. Kontinuerlig feedback**
Giv øjeblikkelig og løbende feedback under hele kurset. Det hjælper med at få deltagernes misforståelser rettet tidligt og styrker deres læring, efterhånden som de gør fremskridt.

**2. Praktiske opgaver**
Formulér opgaver, der kræver, at deltagerne anvender det, de har lært, i praktiske scenarier. De kan f.eks. optimere en eksisterende algoritme eller identificere ineffektivitet i en given algoritme.

**3. Peer review**
Indarbejd *peer review*-sessioner, hvor deltagerne evaluerer hinandens arbejde. Det giver ikke kun mere feedback, men opmuntrer også til fælles læring og kritisk tænkning.

**4. Brug rubrikker**
Udvikl klare, kriteriebaserede rubrikker (eller vejledninger) til retning af opgaver. Det sikrer ensartethed i bedømmelsen og klare forventninger til deltagerne.

**5. Selvvurdering**
Opfordr deltagerne til at vurdere deres eget arbejde ud fra givne kriterier. Det fremmer selvrefleksion og dybere læring.

Disse strategier vil hjælpe undervisere med effektivt at formidle komplekst indhold om algoritmer og samtidig sikre, at deltagerne er engagerede og i stand til at anvende deres viden i praksis.

## 3. At forstå algoritmernes rolle i forskellige sektorer og de potentielle faldgruber, der opstår ved deres brug

At få en dyb forståelse af **algoritmernes rolle på tværs af forskellige sektorer** og samtidig kunne se de potentielle faldgruber er afgørende for at udvikle et omfattende og holistisk syn på de konsekvenser og det ansvar, der er knyttet til anvendelsen af algoritmer. Dette understreger nødvendigheden af grundig indsigt i, hvordan algoritmer kan påvirke forskellige aspekter af samfund, økonomi og ledelse på flere planer. De forskellige anvendelser af algoritmer, lige fra forbedring af virksomhedsdrift til personalisering af brugeroplevelser, viser deres betydelige indflydelse på vores moderne livsform (Mourtzis et al., 2022). Denne indflydelse medfører dog også betydelige risici, såsom potentialet **for bias**, brud på **privatlivets fred** og forstærkning af **eksisterende samfundsmæssige uligheder** (Patel, 2024).

Anvendelsen af algoritmer skal være baseret på etiske overvejelser og **lovgivningsmæssige rammer for** at mindske disse risici (Tsamados et al., 2021). Det er afgørende at udvikle robuste systemer for at sikre opretholdelse af ansvarlighed, så algoritmer ikke viderefører eller forværrer uhensigtsmæssige fremgangsmåder eller bestemte former for uligheder. Dette omfatter strenge test- og valideringsprocesser, hvor algoritmer vurderes, om de overholder princippet om **retfærdighed** og nøjagtighed, især inden for kritiske områder som sundhedssektoren, strafferet og finansverdenen. Interessenter, herunder teknikere, politiske beslutningstagere og offentligheden, skal indgå i løbende dialoger for at undersøge konsekvenserne af algoritmiske beslutninger og sikre, at disse værktøjer tjener samfundets bredere interesser (Donovan et al., 2018).

Desuden er **gennemsigtighed** i algoritmiske processer og beslutninger afgørende. Offentligheden har brug for sikkerhed for, at algoritmiske beslutninger træffes ud fra etiske overvejelser, og deres funktion frit kan kontrolleres. Denne gennemsigtighed er ikke kun afgørende for at opbygge tillid til systemerne, men også for at gøre det muligt for eksperter og tilsynsmyndigheder at udføre revisioner og vurderinger effektivt (Felzmann et al., 2019). Kurser og andre uddannelsesaktiviteter, der øger den algoritmiske viden blandt offentligheden og politiske beslutningstagere, er også afgørende (Reisdorf & Blank, 2021). En sådan indsats vil give flere personer mulighed for at deltage i diskussioner og træffe informerede beslutninger om brugen og reguleringen af disse teknologier.

Forskning og udvikling inden for algoritmisk retfærdighed og etik skal fortsat udvikle sig. Efterhånden som teknologien når til nye stadier, vil der opstå nye udfordringer, som kræver innovative løsninger, der tager højde for både tekniske og etiske spørgsmål. Samarbejde mellem den akademiske verden, industrien og

myndighederne kan gøre det lettere at dele best practice og fremme udviklingen af mere retfærdige og ansvarlige algoritmer.

Diskursen om algoritmer skal også omfatte potentialet for positive forandringer. Når algoritmer designes og implementeres på en ansvarlig måde, kan de forbedre tilgængeligheden, effektiviteten og retfærdigheden. Inden for uddannelse kan algoritmer f.eks. hjælpe med at skabe personlige læringsoplevelser, der tilpasser sig de studerendes individuelle behov, hvilket potentielt kan udfylde huller i uddannelsesniveauet. I sundhedsvæsenet kan prædiktive algoritmer forbedre diagnostik og patientresultater ved at identificere risici og anbefale personlige behandlingsplaner.

Konklusionen er, at integrationen af algoritmer i forskellige sektorer giver både muligheder og udfordringer (Dwivedi et al., 2021). Ved at fremme et miljø, der tilskynder til en praksis, der er etisk, gennemsigtig og inkluderende, kan samfundet udnytte fordelene ved algoritmer og samtidig minimere risiciene ved algoritmerne. Som Williamson (2018) nævner, er en nuanceret forståelse af disse dynamikker afgørende for at realisere det fulde potentiale i algoritmiske applikationer på en måde, der er gavnlig og retfærdig for alle i samfundet.

## 3.1. Algoritmernes rolle i forskellige sektorer



Som Parker et al. (2016) fremhæver, har algoritmer gennemsyret alle sektorer og revolutioneret processer og beslutningstagning. I dette afsnit ser vi nærmere på nogle særlige sektorer, hvor den algoritmiske indflydelse er specielt markant:

**a. Finansverdenen:** AI i finansverdenen har udviklet sig betydeligt fra sin spæde begyndelse til at blive et uundværligt værktøj i branchen. Brugen af AI i finanssektoren kan spores tilbage til 1980'erne, hvor basale computerprogrammer blev tilrettet til at kunne automatisere simple opgaver som dataindtastning og regnskab. Men den virkelige transformationsfase begyndte i slutningen af 1990'erne og begyndelsen af 00'erne med fremkomsten af mere sofistikerede maskinlæringsalgoritmer og eksplosionen i adgang til data (Aksoy & Gurol, 2021). Finansielle institutioner så hurtigt potentialet i AI til at levere mere nøjagtige risikovurderinger, hurtigere transaktionsbehandling og forbedret kundeservice. Dette blev yderligere forstærket af den øgede computerkraft og udviklingen af avancerede dataanalyseteknikker. I 2010'erne begyndte AI at ændre handelsmåder, forsikring, afsløring af svindel og styring af kunderelationer markant (Davenport & Mittal, 2023). Her er nogle eksempler på brug af AI i den finansielle sektor:

- Algoritmehandel: Et af de mest berømte eksempler på vellykket algoritmehandel er Renaissance Technologies, en hedgefond, der er kendt for sin Medallion Fund. Denne fond bruger komplekse matematiske modeller til at analysere og udføre højfrekvenshandler. Ved at anvende

avancerede algoritmer har Renaissance været i stand til at opnå enestående afkast, der langt overgår markedets resultater (Qin, 2012).

- Risikostyring: JPMorgan Chase bruger algoritmer til at styre kreditrisikoen. Deres system vurderer risikoen for misligholdelse af lån baseret på forskellige faktorer, herunder økonomiske forhold, sektorens resultater og individuel kredithistorik. Denne algoritmiske tilgang giver dem mulighed for at skræddersy deres lånetilbud og minimere potentielle tab på uerholdelige fordringer.
- Registrering af svindel: PayPal anvender maskinlæringsalgoritmer til at opdage svigagtige transaktioner. Disse algoritmer analyserer tusindvis af transaktionsattributter i realtid, herunder beløbet, den anvendte enhed og brugerens transaktionshistorik. Ved at identificere mønstre, der afviger fra normen, kan PayPals systemer markere potentielt svigagtige transaktioner med stor nøjagtighed og derved forhindre økonomisk tab og beskytte brugerne.
- Kundeservice: Bank of America bruger en AI-drevet virtuel assistent ved navn Erica. Dette algoritmiske værktøj hjælper kunderne med at administrere deres konti, holde øje med udgifter og foretage betalinger. Erica kan også give opdateringer i realtid om kontoændringer og foreslå måder at spare penge på baseret på forbrugsmønstre, der er analyseret af algoritmer.



**b. Sundhedsvæsenet: AI i sundhedsvæsenet** har oplevet en bemærkelsesværdig udvikling fra tidlig eksperimentel anvendelse til at blive en vigtig komponent i moderne medicinsk praksis. Denne transformation er i høj grad blevet drevet frem af fremskridt inden for maskinlæring, big data-analyse og den stigende digitalisering af patientjournaler. Vi vil her se nærmere på de forskellige facetter af AI-implementering i sundhedsvæsenet, herunder hjælp til diagnosticering, personlig behandling, sygdomsforudsigelse, patientstyring og robotoperationer (Bohr & Memarzadeh, 2020). Her er nogle bemærkelsesværdige eksempler:

- **Hjælp til diagnosticering:** AI's evne til at analysere store datasæt har revolutioneret de diagnostiske processer i sundhedsvæsenet. Ved at integrere AI med billeddannelsesteknologier kan læger opdage sygdomme med større nøjagtighed og hastighed end med traditionelle metoder. IBM Watson Health har vist AI's evne til at forbedre den diagnostiske nøjagtighed. Watsons avancerede AI-algoritmer kan analysere betydningen og konteksten af strukturerede og ustrukturerede data i kliniske notater og rapporter. Den er f.eks. blevet brugt inden for onkologi til at identificere behandlingsmuligheder for kræftpatienter ved at lave krydsreferencer mellem millioner af onkologiske kliniske notater og patientjournaler samt eksisterende litteratur.
- **Personlig behandling:** AI muliggør individuelt tilpasset medicinering ved at se på en persons individuelle genetiske profil, miljøfaktorer og livsstil for at skræddersy en behandling Denne tilgang forbedrer patienternes

resultater betydeligt ved at målrette de behandlinger, der med størst sandsynlighed virker for specifikke patienter. **Tempus** bruger AI til at indsamle og analysere store mængder genomiske og kliniske data for at hjælpe læger med at skabe personlige behandlingsplaner for kræftpatienter. Platformen bruger maskinlæringsalgoritmer til at forstå molekylære og terapeutiske data og forudsige, hvilke behandlinger der sandsynligvis vil være mest effektive for den enkelte patient.

- **Forudsigelse og håndtering af sygdomme**: AI-modeller bruges i stigende grad til at forudsige sandsynligheden for sygdomsudvikling, progression og potentielle komplikationer. Denne proaktive tilgang giver mulighed for tidligere indgreb, hvilket kan redde liv og reducere sundhedsudgifterne. Googles DeepMind har udviklet AI-systemer, der kan forudsige akut nyreskade op til 48 timer, før det sker, med bemærkelsesværdig nøjagtighed. AI-modellen analyserer en bred vifte af sundhedsdata i realtid, hvilket muliggør rettidige indgreb, der kan forhindre forværring og forbedre sundhedsresultaterne.
- **Patienthåndtering og -overvågning**: AI forbedrer patienthåndteringen ved løbende at overvåge sundhedsdata gennem kropsbåren teknologi og IoT-enheder. Disse værktøjer advarer sundhedsaktører om ændringer i en patients tilstand, hvilket giver mulighed for at reagere øjeblikkeligt og løbende justere plejen. Virta Health bruger AI-drevne værktøjer til at håndtere kroniske sygdomme som f.eks. type 2-diabetes. Platformen overvåger patientdata i realtid og justerer behandlingsplaner dynamisk, hvilket hjælper med at holde et optimalt blodsukkerniveau og reducere afhængigheden af medicin.
- **Robotkirurgi**: Robotkirurgi, assisteret af AI, giver høj præcision, reduceret lidelse og hurtigere heling. AI-algoritmer giver kirurgerne data i realtid og kan endda automatisere visse faser af operationen for at forbedre sikkerheden og resultatet. Det robotkirurgiske system, da Vinci, fra firmaet Intuitive bruger AI til at forbedre den kirurgiske præcision. Systemet giver kirurgen et meget forstørret 3D-billede i høj opløsning af operationsstedet og konverterer kirurgens håndbevægelser til mindre og mere præcise bevægelser af bittesmå instrumenter inde i patientens krop.

I takt med at AI fortsætter med at udvikle sig, vokser dens potentiale til at forandre sundhedsvæsenet eksponentielt. Disse eksempler understreger AI's evne til at forbedre nøjagtigheden af diagnoser, skræddersy behandlinger til den enkelte patient, forudsige sygdomsforløb, håndtere kroniske tilstande effektivt og udføre kirurgiske processer med hidtil uset præcision. Integrationen af AI i sundhedsvæsenet optimerer ikke kun patienternes udbytte, men strømliner også sundhedspersonalets arbejde og baner vejen for mere innovative tilgange til medicin og kirurgi. Den igangværende udvikling og anvendelse af AI-teknologier er meget lovende i forhold til at løse nogle af de mest udfordrende problemer i sundhedsvæsenet i dag.

**c. E-handel:** AI forbedrer både kundeoplevelser og **erhvervslivet** betydeligt, især i detail- og e-handelssektoren. Ved at indsamle omfattende mængder forbrugerdata og integrere dem i maskinlæringsalgoritmer kan detailhandlen

udvikle avancerede personaliserings-, anbefalings- og automatiseringsfunktioner. Disse AI-drevne funktioner er nu standard på tværs af shoppingplatforme, og de udmærker sig ved at forbedre kundeengagementet og strømline driften og dermed øge virksomhedens overskud og ressourceudnyttelse. Her er nogle eksempler på brug af AI i e-handel og detailhandel:

- **Forbedrede reklamestrategier**
  Smartly.io: Udnytter AI til at automatisere og optimere reklamekampagner på sociale medier, hvilket reducerer det manuelle arbejde betydeligt og samtidig forbedrer annoncens effekt og engagement med kunder på tværs af platforme.
  eBay: Anvender AI til at tilbyde personlige købsanbefalinger og kunderådgivning, hvilket øger købernes tilfredshed og fastholdelse.
- **Optimering af bilsalg**
  Cox Automotive: Implementerer AI gennem sit Esntial-værktøj for at strømline bilkøbsprocesser online, herunder estimering af betaling og risikovurdering, hvilket forbedrer den digitale kunderejse.
- **Effektive leveringssystemer**
  Gopuff: Integrerer AI for at optimere leveringsruter og sikre hurtig og effektiv levering af dagligvarer gennem sine mikroleveringscentre.
- **Innovativ produktudvikling**
  Mondelez International: Udnytter AI i sin forsknings- og udviklingssektor til at booste innovation og forbedre omkostningseffektiviteten i produktionen af snacks.
- **Personlige shoppingoplevelser**
  Mirakl: Bruger AI til at skræddersy produktanbefalinger, forbedre kundeengagement og salg på digitale platforme.
  Rue Gilt Groupe: Anvender AI til at forfine produktanbefalinger og forbedre shoppingoplevelsen på sine hjemmesider for luksusvarer.
- **Avanceret lagerstyring og supply chain management**
  IBM Watson: Hjælper detailvirksomheder med at strømline driften og personalisere kundeinteraktioner gennem dataanalyse i realtid.
  Anaplan: Udnytter forudsigende intelligens til at forudsige forretningsresultater og optimere detailstrategier, der påvirker alt fra lagerstyring til kundeopkøb.
- **Registrering af forfalskninger og indholdsmoderation**
  3PM Solutions: Anvender AI til at opdage forfalskede produkter og sikre nøjagtigheden af sælgervurderinger på store onlinemarkedsplatforme.
- **Konversations-AI og kundeengagement**
  Valyant AI: Udvikler AI-drevne samtaleværktøjer til quick service-restauranter, der forbedrer kundernes bestillingsoplevelser ved hjælp af stemmebaserede teknologier.

**d. Transport:** Der er ingen tvivl om, at AI omformer transportbranchen ved at forbedre effektiviteten, sikkerheden og brugeroplevelsen. Her vil vi se nærmere på de forskellige måder, hvorpå AI integreres i transportsystemer, fra offentlig transport og logistik til personlig mobilitet og trafikstyring.

- **Autonome køretøjer**
  **Selvkørende biler**: AI er kernefunktionen i selvkørende biler, så de kan navigere, undgå forhindringer og træffe beslutninger i realtid. Virksomheder som Tesla og Waymo er ledende i udviklingen på dette område og forbedrer trafiksikkerheden og køretøjernes effektivitet betydeligt.
  Droner til levering: Virksomheder som Amazon eksperimenterer med droner, der bruger AI til autonom levering af pakker, hvilket vil kunne reducere leveringstid og omkostninger til last-mile-logistik.
- **Trafikstyring og 'smart cities'**
  Intelligente trafiksystemer: AI bruges til at analysere trafikmønstre og optimere trafiklyssekvenser, hvilket reducerer trængsel og forbedrer trafiksikkerheden. Byer som Barcelona og Singapore bruger AI-systemer til at have et jævnt trafikflow i travle vejkryds.
  Forudsigende vedligeholdelse: AI hjælper med at forudsige, hvornår køretøjer til offentlig transport og infrastruktur (som broer og veje) har brug for vedligeholdelse, hvilket forhindrer nedbrud og forlænger levetiden for offentlige anlægsaktiver.
- **Optimering af offentlig transport**
  Ruteplanlægning: AI-algoritmer optimerer bus- og togplaner baseret på realtidsdata og historikken i trafikmønstre, hvilket maksimerer driftseffektiviteten og minimerer ventetiden for passagererne.
  Kapacitetsstyring: I spidsbelastningsperioder **kan AI-systemer forudsige passagerflowet og justere frekvensen af den offentlige transport i overensstemmelse hermed, hvilket forbedrer pendlernes forhold.**
- **Fragt og logistik**
  Automatiseret oplagring: Virksomheder som Amazon bruger AI-drevne robotter til at strømline lagerdriften og øge hastigheden og nøjagtigheden i plukke- og pakkeprocesser.
  Optimeret ruteplanlægning: AI-værktøjer analyserer adskillige variabler som vejr, trafikforhold og leveringsvinduer for at foreslå de mest effektive ruter til forsendelser, hvilket sparer tid og brændstofomkostninger.
- **Forbedrede sikkerhedsfunktioner**
  Systemer til at undgå kollisioner: AI-drevne sensorer og indbyggede systemer i køretøjer kan identificere potentielle farer og gribe ind med det samme for at undgå ulykker.
  Systemer til overvågning af chauffører: Disse systemer bruger AI til at overvåge chaufførernes årvågenhed og generelle helbred for at forhindre ulykker forårsaget af træthed hos chaufføren eller et akut ildebefindende.
- **Kundeservice og support**
  AI-chatbots: Transportvirksomheder bruger AI-chatbots til at yde rejsende assistance i realtid og svare på forespørgsler om køreplaner, priser og aflysninger, uden at mennesker er indblandet.
  Personlige rejseanbefalinger: AI-algoritmer analyserer brugernes præferencer og tidligere adfærd for at tilbyde skræddersyede rejseforslag, der forbedrer kundeoplevelsen.

**e. Uddannelse:** AI forvandler uddannelsessektoren ved at tilbyde skræddersyede læringsforløb, automatisere administrative opgaver og tilbyde uddannelsesmiljøer, der inviterer til fordybelse. Her ser vi på forskellige anvendelser af AI i uddannelsessektoren, med særlig fokus på hvordan AI kan hjælpe underviserne, forbedre elevernes læring og optimere ledelsen af uddannelserne.

- **Personlig læring**
  Adaptive læringsplatforme: AI-systemer som DreamBox Learning og Knewton tilbyder personlig læring ved at tilpasse sig den enkelte deltagers tempo og læringsstil. Disse platforme vurderer deltagernes vidensniveau og justerer automatisk indholdet, så det passer til den enkeltes læringsbehov.
  AI-lærere og -assistenter: AI-drevne læringssystemer som Carnegie Learning tilbyder feedback og hjælp i realtid til studerende og hjælper dem med at forstå komplekse emner gennem individuelt tilpasset undervisning og øvelser.
- **Udvikling af uddannelsesindhold**
  Automatiseret generering af indhold: AI-værktøjer kan hjælpe med at skabe og tilpasse uddannelsesindhold. Værktøjer som Quillionz bruger AI til at lave quizzer og resuméer af lærebøger og artikler, hvilket forbedrer studiematerialet uden omfattende menneskelig involvering.
  Apps til sprogindlæring: Applikationer som Duolingo bruger AI til at skræddersy sprogundervisningskurser til brugerens færdigheder og niveau, hvilket gør sprogindlæring lettere tilgængelig og mere effektiv.
- **Bedømmelse og feedback**
  Automatiserede bedømmelsessystemer: AI automatiserer bedømmelsen af standardiserede tests og endda stileskrivning, hvilket reducerer undervisernes administrative byrde og giver dem mulighed for at fokusere mere på undervisning og mindre på at bedømme.
  Forudsigende analyse: AI-systemer analyserer data om eleverne for at forudsige hvor de læringsmæssigt og karaktermæssigt er på vej hen, hvilket gør det muligt for undervisere at gribe tidligt ind over for elever i farezonen, således at deres læring kan komme tilbage på rette spor.
- **Administrativ automatisering**
  Indskrivnings- og optagelsesprocesser: AI strømliner administrative opgaver som optagelse og indskrivning ved at automatisere databehandling og beslutningstagning, hvilket reducerer papirarbejdet og forbedrer nøjagtigheden.
  Styring af ressourcer: AI-værktøjer hjælper med at styre skolens ressourcer, planlægning og elevernes fremmøde, hvilket optimerer driften og reducerer de faste omkostninger.
- **Forbedret tilgængelighed**
  Hjælpemiddelteknologi: AI-drevne værktøjer som talegenkendelse og tekst-til-tale-omformere gør undervisningsmaterialer mere tilgængelige for studerende med handicap og sikrer inklusion og lige muligheder for alle elever.

Visuelle og auditive læringshjælpemidler: AI-drevne applikationer som Microsofts Immersive Reader hjælper studerende med ordblindhed ved at tilbyde læsehjælp, hvilket beviser, at teknologi kan afhjælpe læringsudfordringer.

- **Fordybende læringsmiljøer**
  Virtual Reality (VR) og Augmented Reality (AR): AI integreret med VR og AR skaber fordybende læringsoplevelser, der simulerer scenarier fra den virkelige verden og gør komplekse emner som videnskab og historie mere engagerende og forståelige.
  Læringsspil og simulationer: AI-støttede læringsspil tilpasser sig de studerendes respons og skaber et dynamisk læringsmiljø, der stimulerer engagementet og forbedrer indlæringen.

**f. Sikkerheds- og militærsektoren:** AI er i stigende grad afgørende for at forbedre sikkerhedsforanstaltninger og militære operationer. Dette afsnit beskriver, hvordan AI-teknologier integreres i forsvarsstrategier, overvågningssystemer og sikkerhedsprotokoller, hvilket forbedrer effektiviteten, samtidig med at komplekse udfordringer inden for national og global sikkerhed håndteres.

- **Forbedret overvågning og monitorering**
  Automatiserede overvågningssystemer: AI-drevne systemer anvendes til at overvåge kritiske områder ved hjælp af dataanalyse i realtid for at opdage usædvanlige aktiviteter eller trusler. Disse systemer kan skelne mellem normal og mistænkelig adfærd, hvilket reducerer menneskelige fejl og svartider betydeligt.
  Droneovervågning: AI-drevne droner bruges til luftovervågning og tilbyder dermed omfattende rekognoscering over svært tilgængelige områder eller konfliktzoner uden at risikere menneskeliv.
- **Cybersikkerhed og forsvar**
  Registrering af trusler og reaktion på truslerne: AI-systemer analyserer mønstre i netværkstrafikken for at identificere potentielle cybertrusler, som f.eks. malware- og ransomware-angreb, meget hurtigere end mennesker ville kunne. Når trusler er opdaget, kan AI også automatisere reaktionerne på dem eller foreslå afbødningsstrategier.
  Kryptering af data: AI forbedrer krypteringsteknikker, hvilket gør det sværere for uautoriserede enheder at afkode følsomme oplysninger og dermed sikre datatransmissioner i militær- og sikkerhedsnetværk.
- **Autonome forsvarssystemer**
  Robottekniske kampenheder: AI bruges til at betjene autonome eller semi-autonome robotenheder, der kan udføre forskellige militære opgaver, fra rekognoscering til kamproller, hvilket reducerer behovet for 'rigtige' soldater i farlige operationer.
  Missilforsvarssystemer: AI-algoritmer hjælper med at forudsige og opfange indkommende trusler med stor præcision, f.eks. missiler eller ubemandede luftfartøjer (UAV'er), hvilket sikrer proaktive forsvarspositioner.
- **Efterretningsanalyse og støtte i beslutninger**

Dataintegration og -analyse: AI-systemer integrerer og analyserer store mængder efterretninger fra flere kilder og giver et omfattende overblik, der hjælper militærstrateger med at træffe informerede beslutninger hurtigt og præcist.
Simulering og træning: AI-drevne simuleringer og virtual reality-miljøer (VR) tilbyder realistisk træning af sikkerhedspersonale og soldater og forbereder dem på en række forskellige scenarier uden de risici, der er ved reel kamp.

- **Vedligeholdelse og logistik**
  Forudsigende vedligeholdelse: AI-værktøjer forudsiger, hvornår militært udstyr har brug for vedligeholdelse, hvilket hjælper med at forhindre funktionsfejl og forlænger levetiden for værdifulde aktiver.
  Optimering af forsyningskæden: AI optimerer logistik og styring af forsyningskæden og sikrer, at ressourcerne fordeles effektivt og leveres, hvor der er brug for dem, især i komplekse militære operationer.

**g. Medier og fritid:** AI er blevet en transformerende kraft i medie- og fritidsindustrien, der forbedrer indholdsskabelse, personaliserer brugeroplevelser og optimerer driften. I dette afsnit undersøges det, hvordan AI-teknologier integreres i forskellige aspekter af medieproduktion, distribution og fritidsaktiviteter.

- **Oprettelse og styring af indhold**
  Automatiseret journalistik: AI-værktøjer som NLG-software (Natural Language Generation) bruges til at skabe nyhedsartikler og reportager, især til datadrevet indhold som sportsresultater og økonomi-nyheder, hvilket giver mulighed for hurtigere publicering og frigør 'rigtige' journalister til mere undersøgende journalistik.
  Video- og musikproduktion: AI-algoritmer hjælper med redigering, lige fra farvekorrektion i videoer til mixning af musik, hvilket strømliner produktionsprocesserne og muliggør mere kreativ eksperimentering.
- **Personalisering og anbefalingssystemer**
  Anbefalinger til medier: Streamingplatforme som Netflix og Spotify bruger AI til at analysere henholdsvis seer- og lyttevaner og tilbyder personlige indholdsanbefalinger for at holde brugerne engagerede og forbedre deres tilfredshed med udbuddet.
  Målrettet annoncering: AI-drevet analyse af brugerdata hjælper medievirksomheder med at skabe og levere målrettede reklamer, hvilket øger effektiviteten i marketingkampagner.
- **Publikumsinddragelse og interaktive medier**
  Chatbots og virtuelle assistenter: AI-drevne chatbots sørger for interaktion med brugerne i realtid og tilbyder kundesupport, indholdsanbefalinger og interaktive oplevelser i spil- og virtual reality-miljøer (VR).
  Augmented Reality (AR) og VR: AI er en integreret del af udviklingen af fordybende AR- og VR-oplevelser, der forbedrer oplevelsen og interaktiviteten i virtuelle miljøer, der bruges i spil, virtuelle ture og uddannelsesindhold.
- **Operationel effektivitet**

Programmatisk markedsføring: AI automatiserer køb og salg af annonceplads og optimerer placering og prisfastsættelse i realtid for at maksimere investeringsafkastet for annoncører og medier.
Distribution af indhold: AI-værktøjer hjælper med at optimere distributionsstrategier ved at analysere forbrugsmønstre for at bestemme de bedste kanaler og tidspunkter for udgivelse af indhold med henblik på at maksimere rækkevidde og engagement.

- **Forbedring af kreative processer**
  Manuskriptskrivning og plotudvikling: AI-værktøjer hjælper med at skrive manuskripter ved at foreslå plotudvikling, dialog og karakterinteraktioner baseret på genrekonventioner og publikums præferencer.
  Kunst og grafisk design: AI-drevne værktøjer hjælper kunstnere og designere ved at foreslå forbedringer, generere ideer og automatisere rutineopgaver, hvilket giver større kreativ frihed og mulighed for at eksperimentere.
- **Event-management og planlægning**
  Håndtering af menneskemængder: AI bruges til at planlægge og styre store begivenheder ved at analysere deltagerdata for at optimere stedets indretning, tidsplanlægning og sikkerhedsforanstaltninger.
  Personlige oplevelser: I fritidsparker og ved events tilbyder AI-systemer personaliserede rejseplaner og oplevelser baseret på de besøgendes præferencer og tidligere adfærd, hvilket øger kundetilfredshed og fastholdelse.

## 3.2. Potentielle faldgruber ved brug af algoritmer

På trods af algoritmers mange fordele kan de også være forbundet med betydelige faldgruber, såsom bias og diskrimination, problemer med privatlivets fred, manglende gennemsigtighed og ansvarlighed og en overdreven afhængighed af automatisering (O'Neil, 2016). Hver af disse faldgruber udgør ikke kun etiske og operationelle risici, men kan også underminere offentlighedens tillid til AI-teknologier. Her ser vi nærmere på vi disse kategorier og giver eksempler fra forskellige sektorer for at illustrere de potentielle farer og fejltrin i implementeringen af AI.

**a. Bias og diskrimination:** AI-systemer kan utilsigtet fastholde og forstærke fordomme, hvis de trænes på datasæt, der ikke er repræsentative eller indeholder fejl, der er baseret på fordomme. Dette problem er særligt vigtigt i sektorer som finansverdenen og sundhedssektoren, hvor sådanne fordomme kan føre til uretfærdig behandling af enkeltpersoner.

- Finansverdenen: AI-systemer, der bruges til kreditvurdering, kan afspejle eksisterende racemæssige eller kønsmæssige bias i historiske data. Hvis en model f.eks. trænes på baggrund af tidligere data om godkendelse af lån, og disse data afspejler historiske fordomme mod en bestemt demografisk gruppe, kan modellen fortsætte med at diskriminere den pågældende gruppe.
- Sundhedssektoren: Der har været tilfælde, hvor AI-diagnoseværktøjer har vist uoverensstemmelser i behandlingsanbefalinger baseret på race- eller kønsforskelle. Et bemærkelsesværdigt eksempel er et AI-system, der fejldiagnosticerede visse hudsygdomme hos mørkhudede personer på grund af et datasæt, der overvejende bestod af lysere hudtoner.

**b. Bekymring for privatlivets fred:** De omfattende data, der kræves for at træne og benytte AI-systemer, giver anledning til betydelige bekymringer for privatlivets fred, især for, hvordan data indsamles, lagres og bruges. Det er tydeligt i sektorer som e-handel og uddannelse, hvor persondata i vid udstrækning bruges til at forbedre kundernes og de studerendes oplevelser.

- E-handel: Forhandlere, der bruger AI til at personliggøre shoppingoplevelser, kan indsamle store mængder personlige oplysninger, lige fra købshistorik til placeringsdata i realtid. Det giver anledning til bekymring for risikoen for databrud og uautoriseret brug af personlige oplysninger.
- Uddannelse: I uddannelsessektoren kan AI-systemer, der bruges til at spore de studerendes fremskridt, indsamle følsomme oplysninger om indlæringsvanskeligheder, sundhedsdata og endda adfærdsmønstre. Det medfører risici i forhold til, hvem der har adgang til disse data, og hvordan de kan bruges uden for uddannelsessammenhæng.

**c. Mangel på gennemsigtighed og ansvarlighed:** AI-systemer fungerer ofte som **"sorte bokse**" med beslutningsprocesser, der er uklare eller skjulte for brugere og lovgivere. Denne mangel på gennemsigtighed kan føre til problemer med ansvarlighed, især når beslutninger har betydelige konsekvenser for folks liv.

- Sikkerhed og militær: I militære anvendelser kan AI-drevet beslutningstagning i autonome våben føre til liv eller død. Manglen på gennemsigtighed i, hvordan disse beslutninger træffes, komplicerer de etiske implikationer og giver anledning til betydelige bekymringer om ansvarlighed.
- Sundhedssektoren: Et AI-system, der bruges til patientdiagnoser, giver måske ikke indsigt i sin beslutningsproces, hvilket gør det vanskeligt for læger at forstå, hvordan systemet er nået frem til en bestemt diagnose. Det gør det ikke kun sværere at stole på AI'ens dømmekraft, men komplicerer også ansvaret i tilfælde af fejldiagnoser.

**d. Overdreven afhængighed af automatisering:** Stor afhængighed af AI kan føre til en forringelse af den menneskelige ekspertise, da færdigheder forringes, når AI-systemer overtager beslutningsprocesserne. Denne overdrevne afhængighed kan være særligt problematisk i miljøer, hvor der er meget på spil, som f.eks. i transportsektoren og finansverdenen.

- Transportsektoren: Luftfartsindustrien har oplevet hændelser, hvor overdreven afhængighed af automatiserede systemer har bidraget til ulykker. Piloter er nogle gange alt for afhængige af autopiloten og har måske ikke tilstrækkelig erfaring med manuel flyvning til at kunne tage over i tilfælde af systemfejl.
- Finansverdenen: **Aktiemarkedet har oplevet flere lyn-krak, som til dels kan tilskrives automatiserede handelsalgoritmer**. Overdreven tillid til disse systemer uden tilstrækkeligt menneskeligt tilsyn kan føre til pludselige markedsnedgange baseret på algoritmiske fejl.



## TIPS OG ANBEFALINGER TIL UNDERVISERE

### Tips og anbefalinger til undervisere om at beskæftige sig med algoritmernes rolle på tværs af forskellige sektorer

**1. Brug casestudier**
Start sessionerne med at undersøge casestudier fra den virkelige verden, hvor algoritmer har spillet en afgørende rolle i forskellige sektorer. Det giver en praktisk forståelse af, hvordan abstrakte begreber anvendes i forskellige brancher som finans, sundhed og e-handel.

**2. Fremhæv etiske overvejelser**

Fremhæv de etiske konsekvenser af algoritmebrug, diskuter potentiel bias, bekymringer om privatlivets fred og ansvarlighed. Dette vil fremme kritisk tænkning og etisk bevidsthed blandt deltagerne.

**3. Tilskynd til interaktive diskussioner**
Facilitér gruppediskussioner om konsekvenserne af algoritmeimplementering. Stil spørgsmål som "Hvad kan gå galt med denne algoritme i en virkelig applikation?" eller "Hvordan kan vi mindske potentielle risici?"

**4. Udnyt multimedieressourcer**
Indarbejd videoer, infografik og podcasts, der illustrerer algoritmernes indvirkning. Denne variation i undervisningen kan have en gavnlig effekt for personer med forskellige læringsstile og gøre indholdet engagerende hele tiden.

**5. Simuler algoritmens indvirkning**
Brug simuleringer eller softwareværktøjer, der giver deltagerne mulighed for at se effekten af algoritmeændringer i simulerede miljøer. Denne praktiske tilgang hjælper med at styrke forståelsen af algoritmer gennem praktisk anvendelse af dem.

## Opvarmningsaktiviteter

**1. Mapping af algoritmeroller**
Tildel deltagerne roller, hvor hver person repræsenterer en del af en algoritme i en bestemt sektor. De skal forklare deres rolles indflydelse og potentielle faldgruber, hvilket fremmer en dybere forståelse af algoritmefunktioner og etiske implikationer.

**2. Debat om algoritmeetik**
Faciliter en debat om en kontroversiel brug af algoritmer, f.eks. ansigtsgenkendelse eller forudseende politiarbejde (*predictive policing*). Det opfordrer deltagerne til at se nærmere på og formulere forskellige perspektiver på etiske udfordringer.

**3. Analyse af scenarier**
Præsenter små grupper for scenarier, der involverer brug af algoritmer i forskellige sektorer, og bed dem om at finde potentielle fordele, risici og de etiske beslutninger, som beslutningstagerne skal tage stilling til.

## Tips og anbefalinger til vurdering af deltagernes læring

**1. Analyse af casestudie**
Giv deltagerne til opgave at analysere et casestudie, hvor der anvendes algoritmer i en bestemt sektor. Vurder deres evne til at identificere både fordelene og de etiske overvejelser som beskrevet i casen.

**2. Gruppepræsentationer**
Få deltagerne til at præsentere forskellige aspekter af algoritmers roller og risici i forskellige sektorer. Vurder deres forståelse baseret på analysens dybde og evnen til at beskæftige sig med etiske overvejelser.

**3. Essays**

Bed deltagerne om at skrive essays, hvor de reflekterer over, hvordan en forståelse af algoritmer kan påvirke deres professionelle eller faglige ansvar og etiske hensyn. Vurder deres evne til at integrere kursets indhold med personlige eller hypotetiske professionelle/faglige scenarier.

**4. Quizzer og prøver**
Brug quizzer til at vurdere forståelsen af nøglebegreber som algoritmisk bias og gennemsigtighed. Inkluder scenariebaserede spørgsmål, der kræver kritisk tænkning ud over udenadslære.

**5. Peer fedback**
Inkorporer peer feedback som en del af vurderingsprocessen. Det giver ikke kun flere perspektiver på deltagerens forståelse, men opmuntrer også til at man engagerer sig i læringsmaterialet, når man skal være opponent og/eller fremlægger.

Ved hjælp af disse strategier kan undervisere effektivt uddanne deltagerne i algoritmernes betydelige indflydelse på tværs af sektorer og sikre, at de forstår både de teknologiske aspekter og de bredere etiske konsekvenser.

## 4. Forstå algoritmisk kompleksitet, optimering og begrænsningerne i algoritmisk beslutningstagning



### 4.1. Algoritmisk kompleksitet

Inden for datalogi kan algoritmer enten skabe eller ødelægge succesen for systemer og apps, fordi de skal køre effektivt. Algoritmisk kompleksitet er som en ledestjerne for udviklere og ingeniører. Det handler om at finde ud af, hvor meget computerkraft forskellige algoritmer har brug for, og vi bruger **Big O-notationen** til at gøre det. Men det handler ikke kun om at få tingene til at køre hurtigere; forståelse af algoritmisk kompleksitet hjælper os med at vide, hvor godt vores algoritmer vil fungere, når vi gør vores systemer større, og hvordan vi skal styre de ressourcer, de har brug for.

Algoritmisk kompleksitet handler dybest set om, hvor hurtigt og hvor meget plads en algoritme skal bruge for at udføre en opgave. Tidskompleksitet fortæller os, hvor lang tid det tager for en algoritme at løse et problem baseret på, hvor stort problemet er. Pladskompleksitet handler på den anden side om, hvor meget hukommelse algoritmen bruger, mens den arbejder. Disse målinger hjælper os med at finde ud af, hvor godt en algoritme fungerer, og om den kan håndtere store opgaver uden at blive for langsom.

I datalogiens verden er der noget, der hedder Big O-notation, som er den matematiske måde at tale om, hvor hurtigt eller langsomt algoritmer kører. Det hjælper os med at forstå, hvor meget tid eller plads en algoritme har brug for, når vi giver den flere ting at arbejde med. Så hvis vi siger, at en algoritme er O(n), betyder det, at når vi giver den mere input (lad os kalde det "n"), tager det lineært

længere tid at blive færdig. Men hvis den er O(n^2), så tager det meget længere tid, når du øger 'n', fordi den vokser kvadratisk. Så Big O-notationen er en hurtig måde til at forstå, hvor effektiv en algoritme vil være, når den håndterer en masse data.

Vigtigheden af at analysere algoritmisk kompleksitet omfatter forskellige områder:

- **Forudsigelse af ydeevne:** Når vi analyserer, hvor komplekse algoritmerne er, kan vi forudse, hvordan de vil fungere med forskellige input. Denne evne til at forudsige ydeevne er afgørende for at vælge den mest hensigtsmæssige algoritme til et bestemt problem og for at optimere systemets ydeevne.

- **Vurdering af skalerbarhed:** I miljøer, hvor inputstørrelser kan ændre sig, hjælper forståelse af algoritmisk kompleksitet os med at vurdere, hvor godt algoritmer kan skaleres. Algoritmer med gode kompleksitetskarakteristika fungerer hensigtsmæssigt på tværs af en lang række inputstørrelser, hvilket sikrer skalerbarhed og responsivitet.

- **Håndtering af ressourcer:** Effektiv brug af ressourcer er afgørende i miljøer, hvor ressourcerne er begrænsede, som f.eks. indlejrede systemer og cloud computing. Analyse af algoritmisk kompleksitet kan medvirke til informeret beslutningstagning om, hvordan man tildeler ressourcer, sikrer optimal brug og minimerer spild.

- **Design af algoritmer:** Forståelse af algoritmisk kompleksitet påvirker designprocessen og guider udviklere i retning af algoritmer, hvor der er balance mellem effektivitet og funktionalitet. Ved at vælge algoritmer med gunstige kompleksitetskarakteristika kan udviklere skabe systemer, der fungerer optimalt uden at gå på kompromis med funktionaliteten.



## 4..2. Tidskompleksitet i algoritmer

Overvej det allestedsnærværende problem: sortering. Algoritmer som *bubble sort*, der har en algoritmisk kompleksitet på O(n2), har en tendens til at fungere dårligt, når de håndterer omfattende datasæt, sammenlignet med mere effektive alternativer som *merge sort* eller *quicksort*, der har en algoritmisk kompleksitet på O(n log n). At forstå denne forskel er afgørende for at kunne træffe velinformerede beslutninger, når man vælger algoritmer. Dette fører igen til forbedringer i systemets ydeevne og forbedrer efterfølgende brugeroplevelsen (Karp, 2010).

For at få en tydeligere illustration kan vi forestille os et scenarie, hvor vi skal sortere et datasæt med en milliard poster (n = 1.000.000.000), hvilket kan sammenlignes med at sortere brugerne af et stort socialt netværk.

- **Boblesortering/*bubble sort*:** Med en algoritmisk kompleksitet på O(n2) vil udførelsestiden være astronomisk, når den anvendes på dette kæmpestore

datasæt. Hvis man skulle give et skøn, ville udførelsestiden være proportional med ca. 1018, hvilket resulterer i et svimlende tal. Hvis hver post kræver 1 nanosekund til sortering, vil den samlede tid, der kræves af bubble sort for at behandle en milliard poster, udgøre 1018 nanosekunder, svarende til 11574,07 dage eller ca. 31,6881 år. Denne ineffektive proces vil sandsynligvis føre til betydelige forsinkelser i sorteringen af data, hvilket giver en dårlig brugeroplevelse.

- **Quicksort:** På den anden side viser quicksort sig at være betydeligt mere effektiv med en tidskompleksitet på O(n log n). Når man sorterer et datasæt med en milliard poster, vil udførelsestiden ved hjælp af quicksort være proportional med cirka 1.000.000.000 * log2(1.000.000.000). Hvis hver post kræver 1 nanosekund til sortering, vil den tid, det tager for quicksort at sortere et datasæt med en milliard poster, være ca. 29.897.352.853,986263 nanosekunder, hvilket svarer til 29,8974 sekunder eller ca. 0,00034603 dage. Derfor vil brugerne opleve hurtigere svartider og en generelt mere jævn brugeroplevelse.

Implementering af mere effektive sorteringsalgoritmer, såsom quicksort, illustrerer det transformative potentiale i optimering af computerprocesser (Bentley, 1999).

Hvorfor bruger vi eksemplet med sortering? Organisering af data er afgørende for, at algoritmer kan arbejde effektivt, og det giver et væld af fordele som f.eks. hurtigere databehandling og smidigere brugerinteraktioner. Forestil dig at søge i et rodet rum efter en bestemt genstand; det er tidskrævende og besværligt. På samme måde kræver usorterede data i algoritmer en lineær søgning, hvor hvert element inspiceres, indtil det ønskede er fundet. Denne proces med en kompleksitet på O(n) (som repræsenterer datasættets størrelse) minder om at scanne en bog – side for side – for at finde et ord. Det er en funktionel, men ineffektiv metode, især ved store datasæt.



Forestil dig nu det samme rum ryddet op, med tingene pænt sat på plads. Det bliver nemt at finde det, man har brug for, fordi man ved, hvor man skal lede. På samme måde gør sorterede data det muligt for algoritmer at bruge binær søgning, hvilket er en betydeligt mere effektiv tilgang. Binær søgning deler søgeintervallet i to gentagne gange, indtil man finder det ønskede element, med en tidskompleksitet på O(log n), hvilket bedre kan skaleres til større datasæt. Det svarer til at finde et ord i en ordbog ved at bladre gennem siderne og halvere søge-volumenet for hver gang.

I bund og grund gør sortering af data det muligt for algoritmer at arbejde smartere, ikke hårdere. Det omdanner en besværlig lineær søgning til en lynhurtig binær søgning, hvilket reducerer behandlingstiden og forbedrer den samlede performance. Så næste gang du støder på usorterede data, skal du huske på sorteringens betydning og dens evne til at frigøre algoritmisk effektivitet.

## 4.3. Algoritmisk kompleksitet vedrørende hukommelsen

Når man ser nærmere på algoritmisk kompleksitet vedrørende hukommelse, støder man på et vigtigt aspekt af algoritmeanalyse, der supplerer undersøgelsen af tidskompleksitet. Hukommelse, som er en begrænset ressource i computermiljøer, har stor indflydelse på algoritmernes ydeevne og systemets effektivitet. At se nærmere på begrebet algoritmisk kompleksitet i forbindelse med hukommelse kan lyde komplekst, men lad os dele det op i mindre dele.

Når vi taler om algoritmisk kompleksitet og hukommelse, ser vi dybest set på, hvor meget plads en algoritme har brug for til at løse et problem. Tænk på det som at lave orden i din taske, så det passer til forskellige opgaver og gøremål - nogle opgaver har måske brug for mere plads, mens andre har brug for mindre.

Forestil dig, at du skal sortere en masse kort med tal på. En måde at gøre det på kaldes *merge sort* (fusionssortering). *Merge sort* deler kortene op i mindre grupper, sorterer hver gruppe og sætter dem derefter sammen igen. Metoden merge sort er rigtig god til at sortere hurtigt, men den har også brug for lidt ekstra plads at arbejde på. Denne ekstra plads svarer til at have ekstra borde til at lave orden i dine kort, mens du sorterer dem.

Et andet eksempel er, når vi udregner Fibonacci-tal, som dem, der findes i naturen (som mønsteret af kronblade på en blomst eller spiralen på en muslingeskal). Der er en måde at beregne disse tal på ved hjælp af en metode, der kaldes dynamisk programmering. Denne metode hjælper os med at finde tallene hurtigere, men den har også brug for ekstra plads til at huske de tal, den allerede har beregnet. Det er som at have en notesbog, hvor man skriver tallene ned, efterhånden som man finder dem.

Lad os nu tale om, hvordan vi lagrer information i computerprogrammer. Vi har forskellige værktøjer kaldet datastrukturer, som f.eks. arrays og lister, der hjælper os med at holde styr på tingene. Hvert af disse værktøjer optager en vis mængde plads i computerens hukommelse.

Forestil dig for eksempel, at du laver en liste med dine venners navne. Hvis du bruger en simpel liste, hvor du skriver deres navne på et stykke papir, fylder den ikke meget. Men hvis du bruger noget mere avanceret, f.eks. en tabel med mange kolonner, fylder den måske mere, fordi du gemmer ekstra oplysninger om hver enkelt ven.

Så når vi analyserer algoritmisk kompleksitet i forbindelse med hukommelse, tænker vi dybest set på, hvor meget plads vores algoritmer og datastrukturer har brug for til at udføre deres arbejde effektivt. Ved at forstå dette kan vi træffe bedre valg i programmeringen for at sikre, at vores programmer kører problemfrit og ikke bruger for meget hukommelse. Det er som at lave orden i sin taske på en smart måde, så man kan have alt det med, man har brug for, uden at den bliver for tung.

Kort sagt hjælper det at tænke på hukommelse og algoritmisk kompleksitet os med at skrive bedre programmer, der fungerer godt og ikke lægger beslag på alle computerens ressourcer. Det er som at være en smart organisator for din computer!

Algoritmisk kompleksitet er et vigtigt begreb inden for moderne databehandling. Det giver en struktureret måde til at vurdere og forbedre, hvor effektive algoritmer er. Ved at bruge Big O-notation kan vi forstå de grundlæggende træk ved algoritmer, og hvordan de fungerer med forskellige datamængder. Forståelse af algoritmisk kompleksitet hjælper udviklere og it-teknikere med at skabe systemer, der kan håndtere de konstant skiftende krav til databehandling. Efterhånden som teknologien udvikler sig, vil viden om algoritmisk kompleksitet fortsat være afgørende for at fremme innovation og forbedre, hvad computere kan udføre.

## 4.4. Optimering

Optimering står som det overordnede pejlemærke for innovation og effektivitet, og optimering er også det, der er styrende for algoritmer, så de fungerer med uovertruffen hastighed og præcision. Kernen i optimering er kunsten at forfine algoritmer for at minimere kompleksiteten og maksimere ydeevnen og derved frigøre deres fulde potentiale til at tackle komplekse problemer på en sofistikeret måde.



Forestil dig algoritmer som de motorer, der driver vores digitale verden, og som er drivkraften i alt – lige fra søgemaskiner til logistiksystemer. Ligesom en perfekt indstillet motor fungerer en optimeret algoritme problemfrit og finder hurtigt vej gennem store datasæt og indviklede beregninger. Men hvordan opnår man lige dette via optimering?

Optimering omfatter et utal af strategier, der har til formål at strømline algoritmer. Det indebærer identifikation af flaskehalse, eliminering af overflødige trin og finjustering af processer for at sikre maksimal effektivitet. I bund og grund handler optimering om at finde den optimale balance mellem ressourceudnyttelse og outputkvalitet.

Et af de grundlæggende mål med optimering er at minimere kompleksiteten. Algoritmer støder ofte på udfordringer, når de får til opgave at løse komplekse problemer, der kræver omfattende beregningsressourcer. Ved at optimere algoritmer sigter vi mod at forenkle deres struktur og reducere arbejdsbyrden med at beregne, så algoritmerne bliver mere håndterbare og skalerbare.

Desuden søger optimering at forbedre ydeevnen ved at udnytte innovative tilgange. Det indebærer at se nærmere på banebrydende teknikker som *parallel computing*, heuristiske algoritmer og maskinlæring for at øge effektiviteten. Ved at udnytte disse fremskridt giver optimering algoritmerne mulighed for at levere hurtigere resultater, samtidig med at de bruger færre ressourcer. Tænk f.eks. på optimering af søgealgoritmer, der bruges af internetbrowsere. Gennem

omhyggelig finjustering og algoritmiske forbedringer kan søgemaskiner hurtigt gennemgå store mængder data for at levere relevante resultater på millisekunder. Denne optimering forbedrer ikke kun brugeroplevelsen, men reducerer også serverbelastningen og energiforbruget.

Hvis vi ser nærmere på parallel databehandling (*parallel computing*) har grafikprocessor-enheder (GPU'er) vist sig at være uundværlige komponenter, især i forbindelse med kunstig intelligens (AI) og maskinlæring. Selvom GPU'er er meget brugt, er der stadig mange, der ikke kender til GPU'ernes egentlige, indre funktioner og deres centrale rolle i udviklingen af AI-teknologier.

Kernen i en GPU er et specialiseret elektronisk kredsløb, der er designet til hurtigt at manipulere og ændre hukommelsen for at fremskynde oprettelsen af billeder i en billedbuffer, der er beregnet til output på en skærm. GPU'er blev oprindeligt udviklet til at gengive grafik, men har udviklet sig til stærkt paralleliserede processorer, der er i stand til at udføre tusindvis af beregningsopgaver samtidigt.

I modsætning til traditionelle CPU'er (Central Processing Units), som udmærker sig ved sekventiel databehandling, er GPU'er optimeret til parallel behandling. De består af mange mindre behandlingsenheder kaldet "kerner", som hver især er i stand til at udføre opgaver uafhængigt af hinanden. Denne parallelle arkitektur gør det muligt for GPU'er at tackle beregningstunge opgaver med høj effektivitet. GPU'er er designet til at parallelisere opgaver, der involverer udførelse af den samme operation på flere data-bidder samtidigt. Denne parallelitet gør det muligt for GPU'er at håndtere beregningsintensive opgaver med bemærkelsesværdig effektivitet (for eksempel grafikrendering og billed- og videobehandling, der involverer udførelse af adskillige beregninger for hver pixel, simulering og modellering, *mining* af kryptovaluta og paralleliserede matrixoperationer).



Parallelisme er særligt fordelagtigt i AI-applikationer, hvor opgaverne ofte involverer behandling af store mængder data samtidigt. Opgaver som træning af dybe neurale netværk, billedgenkendelse, behandling af naturligt sprog og dataanalyse har stor gavn af GPU'ernes evne til parallel behandling.

AI er stærkt afhængig af komplekse matematiske operationer, som f.eks. matrixmultiplikationer og *convolutions* (foldning), som er grundlæggende for træning og implementering af maskinlæringsmodeller. Disse operationer involverer manipulation af store matricer og udførelse af adskillige beregninger samtidig, hvilket gør dem til ideelle til parallel eksekvering på GPU'er.

På dette tidspunkt er vi nødt til at introducere AI-algoritmers strømforbrug. Da disse systemer er blevet allestedsnærværende på tværs af forskellige domæner, er det nødvendigt at fokusere på energieffektivitet for at mindske miljøpåvirkningen og driftsomkostningerne. Optimeringsteknikker og algoritmisk kompleksitet spiller en afgørende rolle for at opnå energibesparelser inden for AI. Optimering indebærer perfektionering af algoritmer for maksimal opgaveudførelse med minimale beregningsressourcer. Energiforbruget i AI-struktorer stammer fra opgaver som datamanipulation og modeltræning. En lav

algoritmisk kompleksitet er afgørende for at reducere energiforbruget, da det kræver færre beregninger.

## 4.5. Begrænsninger i algoritmisk beslutningstagning

Algoritmisk beslutningstagning er blevet et værktøj til at automatisere og optimere processer på tværs af forskellige domæner. Men selv om algoritmisk beslutningstagning har et stort potentiale, har det også iboende begrænsninger, som kræver en grundig undersøgelse. Fra finansverdenen til sundhedssektoren, uddannelse og strafferet er der blevet anvendt algoritmer, som lover uovertruffen effektivitet, nøjagtighed og objektivitet. Alligevel findes der et komplekst netværk af begrænsninger, der indskrænker effektiviteten af algoritmisk beslutningstagning (Yeung, 2018).

**Bias og diskrimination:** En af de dominerende og skadelige begrænsninger, der påvirker algoritmisk beslutningstagning, er tilbøjeligheden til bias og diskrimination. På trods af bestræbelser på at være upartiske, arver algoritmer ofte fordomme i de træningsdata, de bruger. Tilstedeværelsen af historiske uretfærdigheder og samfundsmæssige fordomme i disse data kan opretholde og intensivere forskelle, hvilket resulterer i diskriminerende resultater. Derudover gør uigennemsigtigheden af algoritmiske metoder det vanskeligt at identificere og afbøde bias, hvilket øger bekymringen for retfærdighed og lighed i beslutningsprocesserne. Dette kan udgøre en positiv mulighed for at opdage bias og diskrimination, der tidligere er gået ubemærket hen. Ved at analysere store datasæt ved hjælp af avancerede algoritmer kan man identificere subtile mønstre og forskelle, så der kan træffes korrigerende foranstaltninger for at rette op på systemisk bias. Visse algoritmers gennemsigtighed og sporbarhed gør det muligt at undersøge beslutningsprocessen nøje, hvilket giver involverede muligheden for at finde rødderne til eventuelle bias.

**Mangel på kontekstuel forståelse:** Algoritmisk beslutningstagning sker inden for et kontekstuelt simpelt domæne og er helt afhængig af kvantitative målinger og foruddefinerede regler. Denne begrænsning er særlig udtalt i indviklede og dynamiske miljøer, hvor kontekstuelle finesser er en integreret del af beslutningstagningen. Menneskelig dømmekraft, suppleret med kontekstuel viden og færdigheder inden for området, overgår ofte en algoritmes evner til at skelne mellem nuancerede forviklinger og træffe informerede beslutninger. Som følge heraf kan algoritmiske systemer vise sig at være stive og utilstrækkelige til at operere i skiftende kontekster, og dermed kompromitteres effektiviteten og anvendeligheden af deres beslutninger.

**Uforudsete konsekvenser og etiske dilemmaer:** Algoritmernes deterministiske natur gør dem påvirkelige over for uforudsete konsekvenser og etiske dilemmaer, der stammer fra den iboende oversimplificering og reduktionisme i algoritmiske beslutningsprocesser. Uforudsete resultater, kaskadeeffekter og utilsigtede

konsekvenser kan opstå på grund af algoritmernes manglende evne til at rumme kompleksiteten af scenarier i den virkelige verden. Desuden understreger de etiske dilemmaer, f.eks. vedrørende krænkelse af privatlivets fred, udhuling af menneskelig autonomi og moralsk ansvarlighed, nødvendigheden af omhyggelige overvejelser og tilsyn med implementeringen af algoritmiske systemer.

## 4.6. Hvordan identificerer man potentielle kilder til bias og fejl i algoritmiske systemer? Et afgørende spørgsmål for at sikre retfærdighed, gennemsigtighed og ansvarlighed

En af de primære kilder til bias i algoritmisk beslutningstagning stammer fra de data, der bruges til at træne og teste algoritmerne. Bias i data kan forplante sig i hele beslutningsprocessen og resultere i 'skæve' eller unøjagtige resultater. For at identificere potentielle bias i data er det nødvendigt at foretage en grundig analyse af dataindsamlingsprocessen for at vurdere eventuelle systematiske bias eller underrepræsentation af visse grupper, at undersøge datasættets demografiske karakteristika for at identificere misforhold eller ubalancer, der kan føre til skæve resultater, og at vurdere den indledende databehandling, såsom datarensning og funktionsudvælgelse, for at identificere eventuelle utilsigtede ændringer eller forvrængninger af dataene.



For at identificere potentielle bias og fejl i design og implementering af algoritmer bør man undersøge de algoritmiske modeller og metoder, vurdere de underliggende antagelser og begrænsninger i algoritmen for at afgøre, om de er i overensstemmelse med etiske og samfundsmæssige normer, undersøge træningsprocessen for at identificere potentielle kilder til *overfitting* eller *underfitting*, som kan føre til unøjagtige eller uretfærdige resultater, undersøge valget af funktioner og variabler, der bruges af algoritmen, for at sikre, at de er relevante, ikke-diskriminerende og repræsentative for befolkningen. Overfitting og underfitting er to almindelige problemer, der opstår, når man træner maskinlæringsmodeller. Overfitting opstår, når en model lærer at opfange støj eller tilfældige udsving i træningsdataene i stedet for de underliggende mønstre eller relationer. Resultatet er, at modellen klarer sig godt på træningsdataene, men modellen kan ikke på samme gode måde håndtere nye, ukendte data. På den anden side opstår underfitting, når en model er for simpel til at indfange den underliggende struktur i dataene, hvilket resulterer i dårlig performance både på træningsdataene og på nye data.

For at identificere potentielle bias og fejl i evaluering og validering bør man teste metoder til at måle algoritmens performance på tværs af forskellige demografiske grupper for at opdage forskelle eller uligheder, udføre sensitivitetsanalyser for at vurdere algoritmens robusthed over for variationer i inputdata eller modelparametre og anmode om feedback fra interessenter,

herunder berørte samfundsgrupper og domæneeksperter, for at identificere potentielle bias eller etiske problemer, der er blevet overset under udviklingen.

Det er også nødvendigt med løbende overvågning og revisionsprocesser for at opdage og afbøde bias, fejl eller utilsigtede konsekvenser.

## 4.7. Håndtering og afbødning af algoritmiske risici og bias

Når der er konstateret algoritmiske risici og bias, skal de håndteres på forskellige måder, hvori der kan indgå tekniske, proceduremæssige og etiske overvejelser. Tekniske indgreb kan f.eks. være forbedring af algoritmer gennem teknikker som at fjerne bias fra algoritmer (*debiasing*), diversificering af træningsdata og indarbejdelse af fairness-metrikker i modelevalueringen. Proceduremæssige indgreb kan f.eks. være implementering af robuste valideringsprotokoller, gennemsigtighedsforanstaltninger og ansvarlighedsmekanismer for at sikre, at algoritmiske beslutninger undersøges og valideres effektivt. Etiske indgreb drejer sig om at fremme en kultur med etisk bevidsthed og ansvar blandt udviklere, politiske beslutningstagere og slutbrugere, understrege de etiske konsekvenser af algoritmisk beslutningstagning og fremme etiske retningslinjer og standarder.

At mindske algoritmiske risici og bias kræver løbende overvågning, evaluering og tilpasning til konstant skiftende udfordringer og kontekster. Kontinuerlig overvågning af algoritmisk performance og effekt gør det muligt for interessenter at opdage og håndtere nye bias og risici i tide. Vurderingsmekanismer som bias-audits, sensitivitetsanalyser og konsekvensanalyser giver indsigt, hvor effektive strategier er til reduktion af risici, og de kan danne grundlag for løbende forbedringer. Derudover fremmer tværfagligt samarbejde og engagement udveksling af viden og best practice på tværs af domæner, hvilket beriger den fælles indsats for at minimere algoritmiske risici og bias.



## TIPS OG ANBEFALINGER TIL UNDERVISERE

### Tips og anbefalinger til undervisere om arbejdet med algoritmisk kompleksitet

**1. Konceptuel introduktion**
Begynd med en enkel forklaring på, hvad algoritmisk kompleksitet er, og brug analogier relateret til hverdagsaktiviteter, der kræver planlægning og ressourcer, såsom at planlægge en begivenhed eller en rejse, til at illustrere begreber som effektivitet og ressourceforvaltning.

**2. Visuelle eksempler**

Brug visuelle hjælpemidler som diagrammer og grafer til at vise eksempler på, hvordan forskellige algoritmer performer under forskellige forhold. Det hjælper deltagerne til visuelt at forstå, hvordan kompleksiteten påvirker deres performance.

**3. Praktiske aktiviteter**
Indarbejd praktiske øvelser, hvor deltagerne kan eksperimentere med forskellige algoritmer for at løse specifikke problemer. Det kan være simple kodningsøvelser eller brug af softwareværktøjer, der simulerer algoritmernes performance.

**4. Diskuter algoritmers anvendelse i den virkelige verden**
Knyt diskussionen til anvendelse i den virkelige verden, hvor algoritmisk kompleksitet spiller en afgørende rolle, f.eks. databehandling i store teknologivirksomheder eller ruteplanlægning i GPS-systemer, for at fremhæve kompleksitetens betydning.

**5. Tilskynd til kollaborativ læring**
Facilitér gruppediskussioner og problemløsningssessioner, hvor eleverne kan diskutere, hvilke algoritmer der er bedst at bruge i forskellige scenarier, hvilket fremmer en dybere forståelse gennem elevernes samarbejde.

## Opvarmningsaktiviteter

**1. Sorteringsudfordring**
Planlæg en praktisk aktivitet, hvor deltagerne manuelt sorterer ting (f.eks. bøger eller spillekort) ved hjælp af forskellige metoder til at efterligne sorteringsalgoritmer. Diskuter, hvordan hver metode relaterer til en algoritmes tids- og pladskompleksitet.

**2. Spil med match af algoritme-kompleksitet**
Udtænk et kortspil, hvor deltagerne skal matche forskellige algoritmer med deres Big O-notationer og eksempler på optimal brug. Denne interaktive tilgang hjælper med at styrke forståelsen af teoretiske begreber gennem leg.

**3. Brainstorm om effektivitet**
Få deltagerne til at lave en liste over dagligdagsopgaver, som de udfører, og som kunne optimeres med algoritmer (f.eks. sortering af e-mails eller organisering af filer). Diskuter, hvordan forståelse af algoritmisk kompleksitet kan føre til, at disse opgaver lettere kan udføres.

## Tips og anbefalinger til vurdering af deltagernes læring

**1. Tjek elevernes forståelse**
Brug hyppige korte quizzer til at vurdere elevernes forståelse af nøglebegreber som Big O-notation og forskelle mellem tids- og pladskompleksitet. Det er med til at sikre, at deltagerne forstår de grundlæggende elementer, før de går videre til sværere emner.

**2. Praktiske kodningstest**
Inkluder eventuelt praktiske kodningstests, hvor deltagerne skal skrive kode til eller finde frem til de mest effektive algoritmer til givne problemer. Dette tester deres evne til at anvende teoretisk viden i praktiske scenarier.

**3. Projektbaseret vurdering**
Tildel eleverne et lille projekt, hvor de skal analysere et systems eller en softwares performance og foreslå forbedringer baseret på algoritmisk kompleksitet. Dette vurderer deres evne til at anvende deres læring på systemer i den virkelige verden.

**4. Gruppepræsentationer**
Lad deltagerne arbejde i grupper for at forberede præsentationer af, hvordan algoritmisk kompleksitet påvirker forskellige sektorer som f.eks. finansverdenen, sundhedssektoren eller teknologi. Dette demonstrerer deres forståelse af og evne til at formidle kompleks information.

**5. Essays**
Bed deltagerne om at skrive essays om, hvordan forståelsen af algoritmisk kompleksitet kan påvirke deres arbejde eller studier. Det hjælper dem med at vurdere deres evne til at integrere og reflektere over den viden, de har fået.

Ved hjælp af disse strategier kan undervisere effektivt formidle komplekst indhold om algoritmisk kompleksitet og samtidig sikre, at deltagerne er engagerede, forstår materialet til bunds og kan anvende deres viden i praksis.

## 5. Referencer


Aksoy, T., & Gurol, B. (2021). Artificial intelligence in computer-aided auditing techniques and technologies (CAATTs) and an application proposal for auditors. In *Auditing Ecosystem and Strategic Accounting in the Digital Era: Global Approaches and New Opportunities* (pp. 361-384). Cham: Springer International Publishing.

Bentley, P. (1999). *Evolutionary design by computers*. Morgan Kaufmann.

Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. In *Artificial Intelligence in healthcare* (pp. 25-60). Academic Press.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to Algorithms*. MIT Press.

Davenport, T. H., & Mittal, N. (2023). *All-in on AI: How smart companies win big with artificial intelligence.* Harvard Business Press.

Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation & governance, 12*(4), 505-523. https://doi.org/10.1111/rego.12158

Donovan, J., Caplan, R., Matthews, J., & Hanson, L. (2018). *Algorithmic accountability: A primer*. https://apo.org.au/node/142131

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management, 57*, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society, 6*(1), 2053951719860542. https://doi.org/10.1177/2053951719860542

Karp, R.M. (2010). Reducibility Among Combinatorial Problems. In: Jünger, M., et al. *50 Years of Integer Programming 1958-2008*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-68279-0_8

Knuth, D. E. (1997). *The Art of Computer Programming: Fundamental Algorithms, volume 1.* Addison-Wesley Professional.

Mourtzis, D., Angelopoulos, J., & Panopoulos, N. (2022). A Literature Review of the Challenges and Opportunities of the Transition from Industry 4.0 to Society 5.0. Energies, 15(17), 6276. https://doi.org/10.3390/en15176276

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown.

Parker, G. G., Van Alstyne, M. W., & Choudary, S. P. (2016). *Platform revolution: How networked markets are transforming the economy and how to make them work for you.* WW Norton & Company.

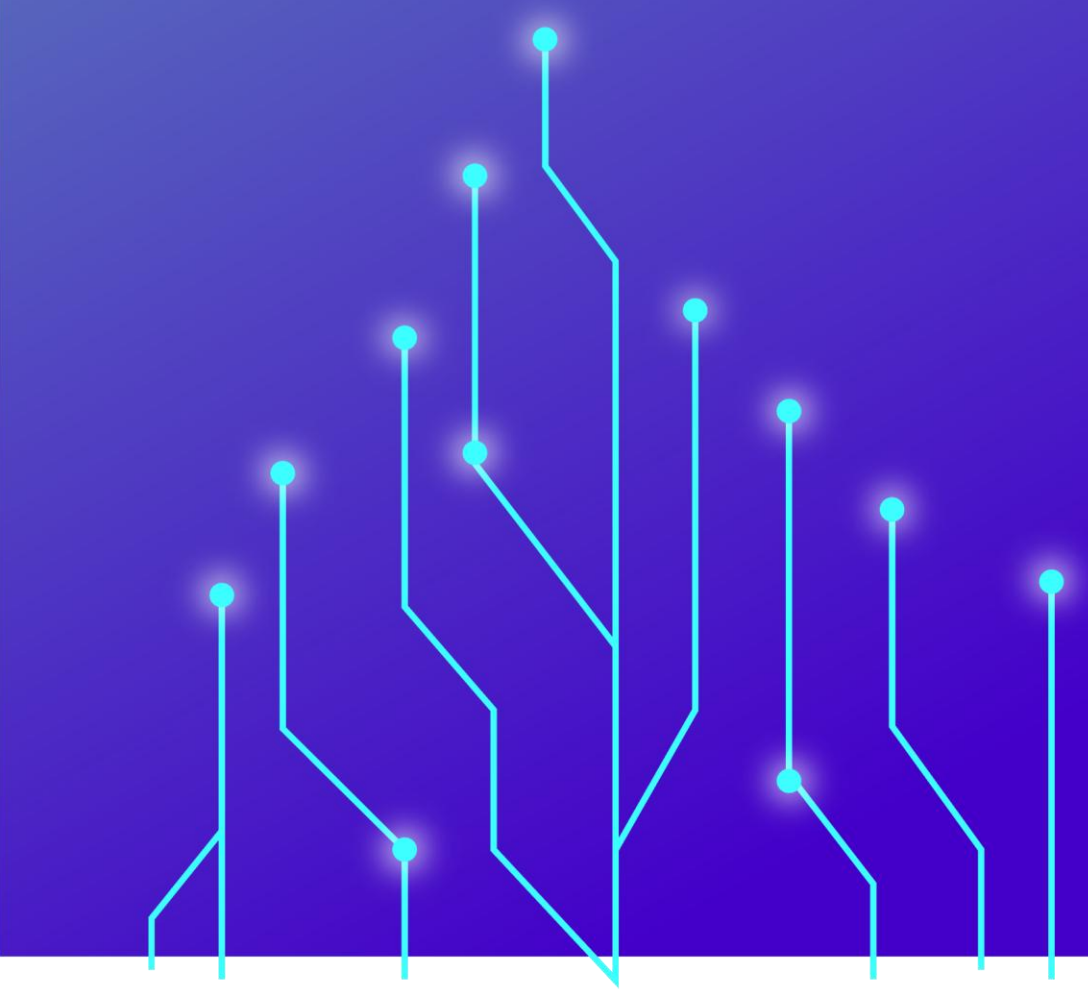
Patel, K. (2024). Ethical reflections on data-centric AI: balancing benefits and risks. *International Journal of Artificial Intelligence Research and Development, 2*(1), 1-17. https://orcid.org/0009-0005-9197-2765

Qin, X. (2012). Making use of the big data: next generation of algorithm trading. In *Artificial Intelligence and Computational Intelligence: 4th International Conference*, AICI 2012, Chengdu, China, October 26-28, 2012. Proceedings 4 (pp. 34-41). Springer Berlin Heidelberg.

Reisdorf, B. C., & Blank, G. (2021). Algorithmic literacy and platform trust. In *Handbook of digital inequality* (pp. 341-357). Edward Elgar Publishing.

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2021). The Ethics of Algorithms: Key Problems and Solutions. In: Floridi, L. (eds) Ethics, Governance, and Policies in Artificial Intelligence. Philosophical Studies Series, vol 144. Springer, Cham. https://doi.org/10.1007/978-3-030-81907-1_8

# CU4 | Retfærdighed og bias i data



# Indhold

Comentado [KM16]: Must be updated in the very end (headlines and page numbers).

# 1. Introduktion

Systemer med **kunstig intelligens** (AI) er ikke bare abstrakte begreber, men er i stigende grad en integreret del af vores hverdag. De påvirker beslutninger i forskellige sektorer – lige fra sundhed og uddannelse til økonomi og sikkerhed. Men i takt med at AI's indflydelse vokser, vokser også potentialet for, at disse systemer kan fastholde eksisterende samfundsmæssige fordomme (bias) eller introducere nye. At forstå begreberne retfærdighed og bias i AI er ikke bare en teoretisk, akademisk øvelse, men det er afgørende for at udvikle retfærdige og upartiske systemer – de systemer, der i stigende grad påvirker vores tilværelse.

Denne kompetenceenhed (*competence unit* – CU) om retfærdighed og bias i AI handler om at forstå dette problem og give dig mulighed for at være en del af løsningen. I kompetenceenheden forsøger vi at belyse, hvordan bias utilsigtet kan opstå i AI-algoritmer, de potentielle virkninger af disse bias og strategier til at afbøde dem. I kurset ses der også nærmere på de etiske, samfundsmæssige og tekniske udfordringer ved at skabe retfærdige AI-systemer, hvor din rolle i etisk AI-udvikling sættes i fokus. Når du har gennemført denne CU, vil du være i stand til at:

**Identificere og forstå forskellige typer af bias**

- Forstå de forskellige typer bias, der kan forekomme i AI-systemer, herunder data-bias, algoritmisk bias og resultat-bias, og forstå de mekanismer, hvorved disse bias manifesterer sig.

**Måle og kvantificere bias**

- Få forståelse for værktøjer og metoder til at vurdere og måle effekten af bias i AI-systemer ved hjælp af en række retfærdighedsmålinger og analyseteknikker.

**Udvikle afhjælpningsstrategier**

- Tilegne sig praktisk viden om, hvordan man implementerer strategier til at håndtere og reducere bias i AI-systemer, herunder forskellige dataindsamlingspraksisser, algoritmiske justeringer og overvågning af implementeringsfasen.

**Vurdere AI-systemers retfærdighed**

- Være i stand til kritisk at vurdere AI-systemers retfærdighed på tværs af flere dimensioner og sikre, at systemerne ikke fastholder eller forværrer sociale uligheder.

**Være beredt til fremtidige udfordringer**

- Kunne forudse og reagere på nye udfordringer inden for retfærdighed i AI, efterhånden som teknologien udvikler sig, og sikre, at retfærdig AI forbliver tilpasningsdygtig og fremsynet.

## TIPS OG ANBEFALINGER TIL UNDERVISERE

Forslag: Begynd med en kort aktivitet, hvor deltagerne deler deres opfattelser af eller personlige erfaringer med AI i deres dagligdag. Det kan hjælpe deltagerne med at forbinde det abstrakte begreb AI med dets praktiske konsekvenser.

### 2. Retfærdighed og partiskhed – "slikanalogien"


BILLEDKILDE | Genereret af DALL-E

Forestil dig, at du sidder i et klasseværelse og kigger på en skål med dit yndlingsslik på lærerens bord. Alle elever må vælge ét stykke, men der er bare det specielle, at elever i røde bluser må få to stykker, mens alle andre må få ét. Det virker ikke fair. Det kaldes "bias" – når nogen (eller noget, f.eks. en AI) giver uretfærdige fordele til bestemte personer uden nogen god grund.

Forestil dig nu, at læreren i stedet beder alle om at løse et puslespil, og at den, der løser det rigtigt, uanset farven på blusen, får et ekstra stykke slik. Det føles retfærdigt, fordi alle har samme betingelser – det handler om puslespillet, ikke om blusen. Retfærdighed er, at alle har lige muligheder, og ingen favoriseres af tilfældige årsager.

Ligesom i slikscenariet bør AI-systemer i den virkelige verden give alle en fair chance, uanset om de anbefaler film, godkender lån eller spiller musik. Når AI er fair, får alle, som lægger puslespillet rigtigt, noget slik.

Når AI er partisk eller biased, er det ligesom, når nogle mennesker får mere slik bare på grund af farven på deres bluse. Og vi vil stræbe efter retfærdighed!

## TIPS OG ANBEFALINGER TIL UNDERVISERE

Forslag: Brug denne analogi interaktivt ved at få deltagerne til at simulere scenariet med ægte slik eller en virtuel aktivitet. Denne praktiske tilgang kan hjælpe med at styrke forståelsen af begreberne retfærdighed og bias.

### 3. At forstå AI-retfærdighed og bias

Vedrørende kunstig intelligens refererer retfærdighed i data til princippet om, at AI-systemer skal behandle alle individer retfærdigt og uden at komme med uretfærdige eller forudindtagede resultater baseret på karakteristika, som er iboende, eller som er kommet til senere. Retfærdighed i AI sikrer, at de beslutninger, der træffes af disse systemer, ikke favoriserer eller forfordeler nogen bestemt gruppe af brugere frem for andre. Det er især vigtigt i applikationer, der påvirker menneskers liv, som f.eks. jobansættelse, godkendelse af lån, retshåndhævelse og sundhedsdiagnostik.



Bias i AI henviser i mellemtiden til systematiske fejl i data eller algoritmer, der fører til uretfærdige resultater for bestemte grupper. Bias kan være eksplicitte, som f.eks. dem, der med vilje er indbygget i et system, men som oftest er de implicitte og skyldes ikke-repræsentative eller ufuldstændige træningsdata eller et fejlbehæftet algoritmisk design, der utilsigtet giver fordele til eller ulemper for bestemte befolkningsgrupper.

BILLEDKILDE | Genereret af DALL-E

Det kan ikke understreges nok, at det er vigtigt at tage fat på retfærdighed og bias i AI-udvikling. AI-systemer, der indeholder bias, kan fastholde og forstærke eksisterende sociale uligheder, hvilket fører til onde cirkler, der er svære at bryde. Desuden nedbryder forudindtagede systemer tilliden til teknologierne, hvilket potentielt kan bremse indførelse af ny teknologi og innovation. At sikre retfærdighed i AI er en teknisk nødvendighed og et moralsk 'must' for at opretholde samfundets tillid til AI og forhindre skadelige virkninger.

## TIPS OG ANBEFALINGER TIL UNDERVISERE

Forslag: Brug casestudier eller nyhedsartikler, der fokuserer på nylige problemer med bias i AI-systemer, og diskutér dem efterfølgende i grupper. Det kan sætte vigtigheden af retfærdighed i AI ind i en kontekst.

### Kønsdiskriminering i jobrekruttering med brug af AI

I 2018 blev det afsløret, at et AI-system, der blev brugt af Amazon til rekruttering, var biased med hensyn til kvinder. AI'en havde lært sig selv, at mandlige kandidater var at foretrække ved at observere mønstre i indsendte CV'er over en 10-årig periode, overvejende fra mænd, hvilket hovedsagelig skete på grund af teknologibranchens ubalance mellem kønnene. Systemet nedprioriterede CV'er, der indeholdt ordet "kvinder", som i "leder af skakklub for kvinder" eller "college for kvinder".



# 4. Bias i AI-systemer

Bias i AI refererer til systematiske fejl, der resulterer i uretfærdige resultater for visse grupper, typisk ved at favorisere en demografisk gruppe frem for andre. Disse bias kan opstå fra forskellige kilder, f.eks. de data, der bruges til at træne AI-systemer, de algoritmer, der behandler disse data, eller bias kan stamme fra dem, der designer og implementerer disse systemer.

## 4.1. Forekomst af bias i AI-systemer

- **Data-bias**: Denne type bias opstår, når de datasæt, der bruges til at træne AI-algoritmer, ikke repræsenterer den bredere befolkning. Det kan ske på grund af historiske uligheder, eller simpelthen fordi dataindsamlingsmetoderne fanger én demografisk gruppe mere grundigt end andre.
- **Algoritme-bias**: Nogle gange kan algoritmerne være designet på måder, der i sagens natur favoriserer bestemte resultater. Det kan skyldes,

hvordan algoritmerne fortolker data, eller de mål, der er sat under udviklingen af AI.

- **Feedback-loops**: Bias kan også opretholdes og forstærkes af feedbackloops, hvor nogle indledende, eller grundlæggende biased beslutninger fører til ny dataindsamling, der forstærker disse fordomme og skaber en cyklus af diskrimination.

## 4.2. Konsekvenser af bias i AI

- **Sociale konsekvenser**: AI-bias kan forstærke eksisterende sociale uligheder og stereotyper og føre til et samfund, hvor digitale teknologier systematisk forfordeler visse grupper. Denne svækkelse af tilliden til AI-teknologier kan få bredere konsekvenser for indførelsen og effektiviteten af AI på tværs af forskellige sektorer.
- **Konsekvenser for individet**: På det personlige plan kan bias i AI føre til uretfærdig behandling inden for kritiske områder som ansættelse, retshåndhævelse, sundhedspleje og finansverdenen. Enkeltpersoner kan blive ofre for AI'ens uretfærdige resultater på grund af fejlbehæftede beslutningsprocesser, der er påvirket af biased AI-systemer.

## 4.3. Håndtering af bias i AI

Det er vigtigt at indføre omfattende strategier i hele AI-udviklingen for at mindske bias i AI. Disse omfatter:

- At sørge for, at datakilderne er varierede for at sikre en bredere repræsentation i træningsdatasættene.
- At udvikle og implementere algoritmiske audits for at identificere og korrigere bias.
- At etablere kontinuerlige overvågningssystemer til at opdage og håndtere nye bias i implementerede AI-systemer.

## 4.4. Oversigt over typer af bias

Det er vigtigt, at udviklere, beslutningstagere og brugere, der ønsker at skabe og interagere med fair og ligeværdige teknologier, forstår de forskellige former for bias, der kan forekomme i AI-systemer. I dette afsnit vil se nærmere på forskellige former for bias, suppleret med eksempler fra den virkelige verden, der viser deres indvirkning.

**1. Bias i udvælgelsen**: Bias i udvælgelsen opstår, når de data, der bruges til at træne et AI-system, ikke repræsenterer målpopulationen eller det tiltænkte scenarie, hvilket fører til skæve resultater. For eksempel kan en AI-model, der er trænet til at genkende ansigtstræk ved hjælp af et datasæt, der primært består af yngre personer, muligvis ikke identificere træk hos ældre voksne, da modellens træning ikke omfattede et repræsentativt udsnit af alle aldersgrupper.

**2. Bekræftelsesbias**: Bekræftelsesbias i AI viser sig, når data eller modeldesignernes fortolkning af dem utilsigtet understøtter allerede eksisterende overbevisninger og ignorerer facts, som beviser det modsatte. En AI, der bruges i ansættelsesøjemed, og som er trænet på CV'er, der overvejende kommer fra ét køn, kan fortsætte med at favorisere kandidater af det køn og dermed forstærke den eksisterende ubalance i jobudvælgelsen.

**3. Selektionsbias**: Selektionsbias er en specifik type udvælgelsesbias, hvor de data, der indsamles for at træne AI, ikke repræsenterer den bredere befolkning præcist. Hvis et stemmegenkendelsessystem hovedsageligt trænes på data fra talere med amerikansk accent, kan det have svært ved at forstå accenter fra andre regioner, f.eks. britiske eller australske, fordi det er trænet på et ikke-repræsentativt udvalg af data.

**4. Algoritmisk bias**: Algoritmisk bias opstår, når de algoritmer, der ligger til grund for AI-systemer, giver partiske eller biased resultater, ofte på grund af iboende fejl i deres design eller de målsætninger, der blev fastsat i designfasen. En kreditvurderings-AI, der uforholdsmæssigt ofte afviser lån til ansøgere fra bestemte kvarterer på grund af historiske data, der afspejler tidligere bias i forhold til udlån, er et eksempel på algoritmisk bias.

**5. Bias i rapporteringen**: Bias i rapporteringen opstår, når de data, der indlæses i et AI-system, er skæve på grund af tendensen til kun at rapportere visse typer resultater. For eksempel kan et AI-overvågningssystem for medicinbivirkninger, der primært modtager rapporter om alvorlige tilfælde, fejlagtigt antyde, at de fleste patienter oplever ekstreme reaktioner, og ignorere mildere, men mere almindelige bivirkninger.



**6. Målingsbias**: Målingsbias indebærer fejl i dataindsamlingen, som systematisk skævvrider dataene. En AI til miljøovervågning, der er afhængig af sensorer, som er mindre effektive i koldere temperaturer, vil have et skævt billede af forureningsniveauerne og potentielt underrapportere emissioner og udslip i vintermånederne.

## 5. Metoder til at identificere og måle bias i AI

At forstå og mindske bias i kunstig intelligens er afgørende for at opbygge retfærdige og effektive systemer. I dette afsnit beskrives metoder til at identificere og måle disse bias og giver AI-udviklere og dataforskere en klar vej til at sikre, at deres AI-systemer fungerer uden unfair bias.

- **Datarevision** eller audits er en omfattende gennemgang, hvor datasæt, der bruges til at træne AI-modeller, undersøges for potentielle bias eller anomalier. For at udføre en datarevision skal man:
    - **Vurdere dataenes repræsentativitet**: Vurder, om datasættet nøjagtigt afspejler mangfoldigheden i den befolkning, som modellen skal bruges på. Dette omfatter kontrol af, om datasættet

er demografisk repræsentativt, samt om det er dækkende i forhold til forskellige scenarier.

- **Identificere manglende data**: Se efter mønstre i manglende data – huller i dataene kan medføre skævheder, hvis specifikke undergrupper er underrepræsenterede.
- **Analysere om mærkningen af data er konsistent**: Sørg for, at processen med mærkning af data er konsistent og upartisk. Uoverensstemmelser i mærkningen af data kan føre til betydelige skævheder i modellernes resultater.
- **Bruge statistisk analyse**: Brug statistiske værktøjer til at opdage uregelmæssigheder eller skæve datadistributioner, der kan indikere forudindtagede data.

- **Disparitetsanalyse** indebærer sammenligning af, hvordan en model præsterer på tværs af forskellige demografiske grupper eller andre relevante undergrupper for at identificere uoverensstemmelser, der kan indikere, at der er bias. Trinene til at udføre en disparitetsanalyse omfatter:
  - **Definition af relevante undergrupper:** Baseret på attributter som alder, køn, etnicitet eller andre karakteristika, der er relevante for modellens anvendelse.
  - **Præstationsmålinger:** Beregn præstationsmålinger som nøjagtighed, præcision, *recall* og F1-score for hver undergruppe.
  - **Sammenligning af resultater:** Analyser forskellene i præstationsmålinger mellem grupperne. Betydelige forskelle kan indikere tilstedeværelsen af bias.
  - **Analyse af den underliggende årsag:** Hvis der findes forskelle, skal du undersøge de underliggende årsager, som kan vedrøre dataindsamling, modelarkitektur eller valg af funktioner.
- **Sensitivitetsanalyse** tester en AI-models robusthed ved at ændre inputdata en smule og observere, hvordan output ændrer sig. Det er især nyttigt til at identificere bias:
  - **Varier inputdata**: Lav små ændringer i inputdata, der afspejler plausible variationer i scenarier fra den virkelige verden.
  - **Overvåg udsving i output**: Se, om disse ændringer fører til uforholdsmæssigt store ændringer i modellens output.
  - **Vurder indvirkningen på specifikke grupper**: Fokuser på, hvordan disse ændringer påvirker modeloutput for forskellige demografiske grupper.
  - **Foretag justeringer**: Brug resultaterne til at forfine dataforbehandlingen, funktionsudviklingen eller modellen for at reducere uønsket sensitivitet.

## 6. Strategier til at håndtere og mindske bias

At mindske bias i AI-systemer indebærer forskellige tilgange i design-, udviklings- og implementeringsfasen. Følgende strategier er vigtige for at **reducere risikoen for biased resultater**.

- **Mangfoldig repræsentation:** At sikre mangfoldighed i data og teamsammensætning er afgørende for at udvikle upartiske AI-systemer.
    - **Mangfoldighed i data**: Indsaml data fra en bred vifte af kilder for at dække et bredt spektrum af befolkningen. Dette omfatter forskellige demografiske forhold, socioøkonomiske baggrunde og andre relevante karakteristika.
    - **Teamets mangfoldighed**: Forskellige teams bidrager med forskellige perspektiver, der kan identificere og afbøde potentielle bias, som måske ikke er tydelige for en mere homogen gruppe.
- **Bias-bevidste algoritmer:** At udvikle algoritmer, der generelt er bevidste om og kan justere for bias, er en proaktiv tilgang til at mindske bias.
    - **Indarbejd metoder til måling af retfærdighed**: Integrer hensynet til retfærdighed i algoritmens funktion, som f.eks. at minimere forskelle i fejlrater på tværs af grupper.
    - **Adversarial træning**: Implementer adversariale træningsteknikker, hvor modeller trænes til både at forudsige korrekt og til at minimere bias.
- **Regelmæssig overvågning og evaluering**: Kontinuerlig overvågning og regelmæssig evaluering af AI-systemer er afgørende for at sikre, at de forbliver upartiske i længden og tilpasser sig nye data eller kontekster.
    - **Etabler rammer for overvågning**: Opsæt systemer, der løbende sporer AI-modellernes ydeevne, primært med fokus på fairness-metrikker.
    - **Regelmæssige gennemgange**: Gennemgå og rekalibrer regelmæssigt modellerne baseret på løbende overvågningsresultater, og tilpas dem til ændringer i datamønstre eller samfundsnormer.
    - **Feedback fra interessenter**: Samarbejd med interessenter, herunder slutbrugere, for at indsamle feedback om AI's performance og interessenternes oplevede retfærdighed, som kan give indsigt i det, der ikke opfanges i kvantitative vurderinger.

## 7. At forstå retfærdighedsmålinger

Retfærdighedsmålinger er afgørende for at vurdere, hvor godt AI-systemer behandler individer eller grupper med forskellige baggrunde. De giver en ramme til at måle og sikre, at AI-systemer ikke opretholder eller forværrer eksisterende samfundsmæssige bias. Nedenfor er nogle af de **kritiske retfærdighedsmålinger**, der bør overvejes:

- **Statistisk paritet (demografisk paritet)**: Denne metrik vurderer, om andelen af positive resultater er den samme på tværs af forskellige demografiske grupper (f.eks. køn, race). Dette er afgørende for applikationer, hvor målet er at sikre jævnbyrdige resultater uanset inputfunktioner relateret til gruppeidentitet.
- ***Equalized Odds***: Denne metrik kræver, at den sande positive rate (sensitivitet) og den falske positive rate (specificitet) er den samme på tværs af grupper. Den bruges primært i situationer, hvor konsekvensen af en fejl påvirker alle grupper lige meget, f.eks. i strafferet eller ved ansættelser.
- **Betinget demografisk forskel**: Denne metrik evaluerer forskelle i forudsigelige resultater betinget af legitime egenskaber. Den måler forskelle, der ikke er begrundet i relevante egenskaber, og er velegnet til scenarier, hvor der forventes visse forskelle mellem grupper på grund af disse egenskaber.
- **Retfærdighed gennem bevidsthed**: Denne tilgang indebærer, at man indarbejder bevidsthed om følsomme egenskaber (som race eller køn) direkte i beslutningsprocessen for at sikre retfærdige beslutninger. Denne metrik taler for algoritmer, der kompenserer for historiske uretfærdigheder eller samfundsmæssige fordomme, der kan påvirke retfærdigheden.
- **Individuel retfærdighed:** Til grund for denne metrik ligger, at lignende individer bør modtage lignende forudsigelser uanset deres gruppemedlemskab. Dette koncept er især relevant i personaliserede tjenester som jobanbefalinger, hvor ensartet behandling på tværs af lignende brugerprofiler er afgørende.
- **Kontrafaktisk retfærdighed**: I henhold til denne beregning anses en beslutning for at være retfærdig, hvis den ville være blevet truffet i en kontrafaktisk verden, hvor personen tilhørte en anden demografisk gruppe, men ellers var identisk. Dette mål er værdifuldt til at vurdere retfærdighed på individniveau og hjælper med at forstå specifikke egenskabers indvirkning på beslutninger.

Valg af passende retfærdighedsmetrikker afhænger af den specifikke kontekst og formålene med AI-applikationen. Her er nogle retningslinjer, du kan overveje:

1. **Applikationsspecifikke krav:** Valget af retfærdighedsmetrik skal være i overensstemmelse med formålene og de juridiske krav i applikationsdomænet. For eksempel kan *Equalized Odds* være afgørende for strafferetlige applikationer, mens statistisk paritet måske er mere passende for marketing eller reklame.
2. **Lovmæssige og etiske overvejelser:** Nogle brancher kan have lovkrav, der dikterer specifikke retfærdighedsmålinger. Etiske overvejelser, som f.eks. behovet for at rette op på historiske uretfærdigheder, kan også styre valget af retfærdighedsmålinger.
3. **Vurdering af indvirkningen:** Det er vigtigt at overveje AI-systemets potentielle sociale indvirkning og vælge parametre, der minimerer skadelige bias. For eksempel kan kontrafaktisk retfærdighed være afgørende på områder som sundhedspleje, hvor individualiseret behandling er nødvendig.
4. **Teknisk gennemførlighed:** Afhængigt af AI-systemets kompleksitet og datatilgængelighed kan nogle målinger være mere udfordrende at implementere end andre. Tekniske begrænsninger bør overvejes for at sikre, at de valgte retfærdigheds-mål kan anvendes præcist og konsekvent.

## 8. Retfærdighed i dataindsamling og forbehandling



At sikre, at dataindsamlingspraksis er mangfoldig og ikke-biased, er grundlæggende for at kunne opbygge retfærdige AI-systemer.

- **Repræsentativ dataindsamling**: Sørg for, at dataindsamlingsprocessen omfatter prøver fra alle de befolkningssegmenter, som AI-systemet skal servicere. Det kan indebære stratificeret dataindsamling for at inkludere underrepræsenterede grupper.
- **Løbende dataindsamling**: Opdater og udvid regelmæssigt dataindsamlingen, så den afspejler nye og spirende tendenser i befolkningen, da statiske datasæt hurtigt kan blive forældede.

### 8.1. Teknikker til forbehandling

Det er vigtigt at sikre retfærdighed i dataindsamling og forbehandling for at forhindre, at bias siver ind i AI-systemer. Nedenfor er der flere fremgangsmåder, der kan hjælpe med at opnå et mere afbalanceret og retfærdigt AI-system gennem omhyggelig opmærksomhed på de indledende faser af AI-udvikling.

- **Håndtering af manglende data**: Analysér mønstre af manglende data, da disse kan forårsage bias. Teknikker som 'imputation' bør anvendes varsomt for at undgå yderligere bias.

- **Udvælgelse af funktioner**: Vurdér kritisk, hvilke funktioner der indgår i modeltræningsprocessen. Funktioner, der er korreleret med følsomme egenskaber, skal håndteres forsigtigt for at undgå proxy-diskrimination.
- **Normalisering og standardisering**: Anvend normalisering eller standardisering for at sikre, at modellen ikke som en selvfølge vægter visse funktioner tungere blot på grund af deres numeriske skala.
- **Forskellige datakilder:** Det er afgørende at sikre, at dataindsamlingen omfatter et bredt spektrum af befolkningsdemografi for at forhindre bias i AI-modeller. Målet er at indsamle data, der afspejler variationen og mangfoldigheden i den virkelige verden, hvor AI-systemet skal fungere. Det indebærer at samarbejde med forskellige demografiske grupper og bruge flere indsamlingssteder. Når man f.eks. udvikler en AI til sundhedssektoren, skal der indsamles data fra hospitaler i forskellige regioner, herunder land- og byområder, for at indfange forskellige sundhedsprofiler og miljømæssige påvirkninger på sygdomme.
- **Inkluderende dataindsamling**: Inkluderende dataindsamling har til formål at sikre, at alle befolkningssegmenter er repræsenteret i træningsdatasættet. Målet er at forhindre udelukkelse af minoriteter eller underrepræsenterede grupper, hvis fravær kan føre til biased AI-resultater. Der anvendes teknikker som stratificeret dataindsamling, hvor befolkningen opdeles i undergrupper (strata), der afspejler vigtige karakteristika (f.eks. etnicitet, alder, køn), og data udvælges tilfældigt fra hvert stratum for at sikre inklusivitet i datasættet.
- **Rensning af data**: Datarensning indebærer at fjerne unøjagtigheder og uoverensstemmelser, der kan skævvride AI-modellens resultater. Målet er at sikre, at datakvaliteten er høj og fri for fejl, der kan påvirke modellens beslutninger. Denne proces omfatter korrektion eller fjernelse af outliers, udfyldning af manglende værdier ved hjælp af passende imputeringsstrategier, der ikke skævvrider dataene, og sikring af, at alle dataindtastninger er konsekvente og korrekt formaterede.
- **Regelmæssig overvågning**: Overvågning af data og modellers ydeevne er afgørende for at opdage og håndtere eventuelle bias, der kan opstå over tid. Målet er opretholdelse af løbende overvågning for at sikre, at AI-systemet fortsætter med at fungere retfærdigt, når der indsamles nye data, og når forholdene ændrer sig. Etablering af automatiserede overvågningssystemer, der løbende vurderer AI-output for retfærdighed og advarer udviklere, hvis der opdages skævheder. Dette system bør regelmæssigt evaluere de data, der tilføres AI'en, for at tjekke, om der er ændringer i datakvalitet eller repræsentation.
- **Retfærdig funktionsudvikling:** Retfærdig funktionsudvikling indebærer at vælge og at omdanne funktioner for at minimere bias, samtidig med at dataenes anvendelighed bevares. Målet er at sikre, at modelinput ikke i sig selv er til ulempe for nogen gruppe. Teknikkerne omfatter analyse af

funktionernes sammenhæng med sensitive egenskaber og enten ændring eller fjernelse af funktioner, der kan føre til biased resultater. Teknikker til reduktion af dimensioner kan også identificere og fjerne overflødige eller irrelevante funktioner, der kan medføre bias.

- **Gennemsigtighed og dokumentation:** Opretholdelse af gennemsigtighed gennem detaljeret dokumentation af dataindsamling, forbehandling og funktionsudvælgelsesprocesser. Målet er at opnå klare records, der kan revideres for overholdelse af standarder for retfærdighed. Opbevaring af detaljerede logfiler over beslutninger om datahåndtering, metoder til datarensning og rationalet bag udvælgelse af funktioner. Disse dokumenter skal være let tilgængelige for interessenter og tilsynsmyndigheder, så de kan gennemgå og vurdere, om AI-systemet er retfærdigt.
- **Registrering af bias**: Registrering af bias indebærer, at man identificerer potentielle bias i datasættet, før det bruges til at træne AI-modeller. Målet er at korrigere eventuelle bias, der påvirker modellens retfærdighed, på forhånd. Brug af statistiske analyse- og visualiseringsværktøjer til at undersøge datafordelinger på tværs af forskellige demografiske grupper. Dette kan involvere specialiseret software eller algoritmer, der er designet til at fremhæve områder, hvor data måske ikke er repræsentative, eller hvor resultater kan påvirke visse grupper uforholdsmæssigt meget.



Et AI-systems retfærdighed starter med, hvordan data indsamles, behandles og forberedes, før de overhovedet når frem til den algoritmiske træningsfase. Ved at implementere en streng praksis for datahåndtering og forbehandling kan udviklere reducere risikoen for bias i AI-systemer betydeligt. Denne praksis forbedrer ikke kun retfærdigheden og troværdigheden af AI-applikationer, men opbygger også tillid blandt brugere og interessenter i både forskelligartede og regulerede omgivelser.

## 9. Nye tendenser inden for retfærdighed i AI

Efterhånden som AI-teknologier udvikler sig, konfronteres de med komplekse datasæt og lovgivning, som er i stadig forandring. Nye teknologier som *explainable AI (XAI)* og *federated learning* giver mulighed for at forbedre retfærdigheden i AI ved at skabe større gennemsigtighed i AI-beslutningsprocesser og give mulighed for decentral databehandling, der kan hjælpe med at beskytte den enkeltes privatliv. Fremtidige udfordringer er for eksempel:

- **Komplekse datainteraktioner**: Med fremkomsten af big data bliver det mere udfordrende at forstå interaktioner og sammenhænge i store og komplekse datasæt, men det er afgørende for at sikre retfærdighed.

- **Adaptive regler**: I takt med at offentlighedens bevidsthed om AI-bias øges, vil lovgivende organer sandsynligvis indføre strengere retningslinjer for retfærdighed og kræve, at AI-systemer er tilpasningsdygtige og gennemsigtige med hensyn til fortsat at værne om retfærdighed.
- **Forklarlighed og forståelighed**: Forklarlighed og forståelighed i AI handler om at gøre AI-systemernes drift og beslutninger gennemsigtige og forståelige for mennesker. Målet er at sikre, at interessenter forstår, hvordan AI-beslutninger træffes, hvilket er afgørende for tillid og ansvarlighed. Denne tendens indebærer udvikling af metoder og værktøjer, der giver mulighed for at visualisere og forklare AI-beslutningsprocesser. Teknikker som *feature importance scores*, beslutningstræer og model-agnostiske metoder giver indsigt i komplekse modellers virkemåde, f.eks. dybe neurale netværk.
- **Adversarial-angreb og forsvar**: Adversarial-angreb (eller kontradiktoriske angreb) indebærer manipulation af AI-input for at narre AI-systemer til at træffe forkerte beslutninger. Forsvaret mod sådanne angreb har til formål at gøre AI-systemer mere robuste og sikre og beskytte dem mod ondsindede input, der kan udnytte bias eller svagheder i systemet. Implementering af forsvarsmekanismer mod adversarial-angreb omfatter brug af teknikker som kontradiktorisk træning, hvor modellen udsættes for kontradiktoriske eksempler under træningen for at lære at modstå dem. Andre tilgange er regelmæssige sikkerhedsvurderinger og opdatering af modeller, for at de kan lære at genkende og imødegå nye angrebsstrategier.
- **Retfærdighed på tværs af flere dimensioner:** Retfærdighed på tværs af flere dimensioner vedrører håndtering af bias, der går på tværs af forskellige demografiske og socioøkonomiske forhold, herunder, men ikke begrænset til, race, køn, alder og handicap. Målet er at sikre omfattende retfærdighed, der tager højde for menneskers ofte forskelligartede karaktertræk og deres samspil. Det indebærer udvikling af flerdimensionelle retfærdighedsmålinger og brug af tværgående dataanalyseteknikker til at vurdere og afbøde bias i krydsfeltet mellem flere karaktertræk.
- **Retfærdighed i nye teknologier**: Efterhånden som AI integreres i nye teknologier, såsom selvkørende køretøjer, intelligent sundhedspleje og IoT-enheder, bliver det afgørende at sikre retfærdighed i disse applikationer. Målet er at implementere retfærdighed fra bunden i disse nye teknologier for at forhindre bias, der kan have omfattende samfundsmæssige konsekvenser. Det indebærer at gennemføre konsekvensanalyser og bias-audits for nye teknologier og udforme retningslinjer for retfærdighed, der er specifikke for hver teknologis kontekst og anvendelse.
- **Etiske overvejelser og styring**: Etiske overvejelser og styring inden for AI indebærer etablering af rammer og retningslinjer, der sikrer, at AI-systemer udvikles og implementeres på en måde, der overholder etiske

standarder og samfundsmæssige værdier. Målet er at skabe en regulerende og etisk overvågningsmekanisme, der kan være vejledende for udviklingen og brugen af AI. Det kan f.eks. ske gennem udvikling af omfattende etiske retningslinjer for AI, dannelse af etiske råd og samarbejde med politiske beslutningstagere om at vedtage regler, der håndhæver retfærdighed, gennemsigtighed og ansvarlighed inden for AI.

- **Menneskecentreret design**: Menneskecentreret design skaber AI-systemer, der tager hensyn til menneskers ve og vel, værdier og etik. Målet er at sikre, at AI-teknologier forbedrer menneskets formåen og livskvalitet uden at gå på kompromis med eller mindske menneskeligt input og tilsyn. Det indebærer at inddrage forskellige brugergrupper i designprocessen, at indarbejde feedback-loops, der giver brugerne mulighed for at give feedback til den løbende AI-udvikling, og at sikre, at AI understøtter snarere end erstatter menneskelig beslutningstagning.
- **Algoritmisk afhjælpning af bias:** Algoritmisk afhjælpning af bias indebærer at identificere og korrigere bias i AI-algoritmer for at forhindre uretfærdige resultater. Målet er at forfine AI-modeller løbende for at sikre, at de er så objektive og upartiske som muligt. Dette omfatter brug af teknikker til registrering og korrektion af bias under modeltræningsprocessen, f.eks. genvægtning af træningsdata, justering af modelparametre for at bringe nøjagtighed og retfærdighed i balance samt anvendelse af algoritmer til afhjælpning af ekstern bias.

Disse tendenser og tiltag understreger AI-udviklingens dynamiske karakter og det fortsatte behov for innovation og årvågenhed for at sikre retfærdighed, efterhånden som AI-systemer bliver mere integreret i alle aspekter af livet.



## 10. Konklusion

Det er helt essentielt at se nærmere på AI-retfærdighed og bias, i takt med at vi integrerer kunstig intelligens i flere områder af vores dagligdag og vores kritiske infrastruktur. Det er også tydeligt, at retfærdighed i AI er en kompleks udfordring, der kræver omfattende og løbende overvågning – lige fra at have identificeret forskellige bias i AI-systemer og konsekvenserne af dem til strategierne for at måle og afbøde dem.


BILLEDKILDE | Genereret af DALL-E

- **Forståelse af forskellige bias og deres effekter:** Det første skridt mod afhjælpning er at genkende forskellige typer af bias, såsom udvælgelses-, algoritme- og bekræftelsesbias, og forstå deres potentiale til at opretholde uligheder. Eksempler fra den virkelige verden som unøjagtigheder i ansigtsgenkendelse og biased ansættelsesalgoritmer er eksempler på de skadelige virkninger af ukontrollerede AI-bias.

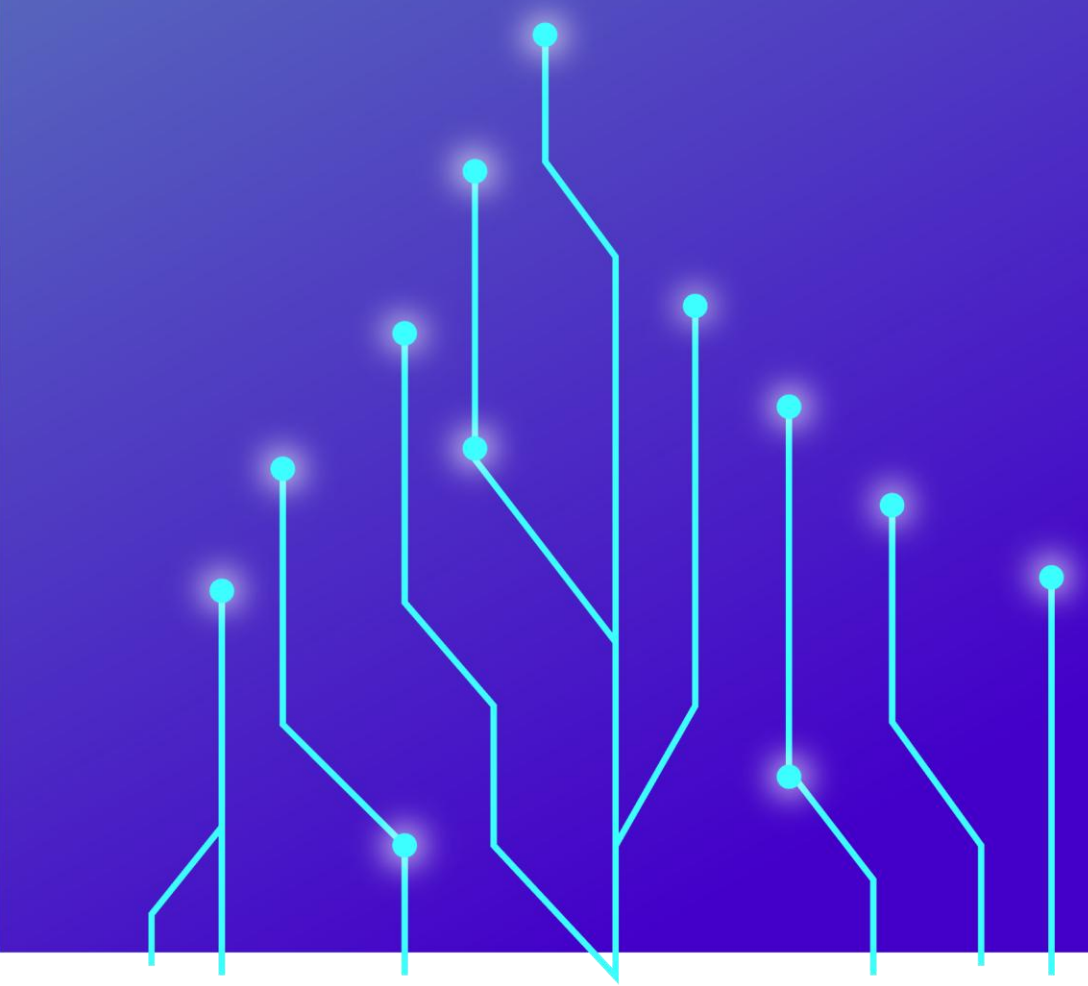
- **Strategier for retfærdighed:** Effektive metoder til at sikre retfærdighed omfatter forskelligartet og inkluderende dataindsamling samt regelmæssig overvågning af AI-systemer. Disse strategier hjælper med at skabe AI-modeller, der er repræsentative og retfærdige, og reducerer systematiske fejl, der forfordeler visse grupper.

- **Nye tendenser og udfordringer:** Retfærdighed i AI er under stadig udvikling, og nye udfordringer opstår, som f.eks. at sikre retfærdighed i nye teknologier og udvikle forsvar mod fjendtlige angreb. Der er også et voksende fokus på forklarlighed og menneskecentreret design, hvilket også er vigtigt for at opbygge troværdige AI-systemer.

- **Etiske overvejelser og styring:** Betydningen af en etisk styring i udviklingen af kunstig intelligens kan ikke overvurderes. Etablering af robuste etiske retningslinjer og styringsrammer vil være afgørende for at styre en ansvarlig udbredelse af AI-teknologier.

- **Kontinuerlig læring og tilpasning:** Når AI-teknologier og samfundsnormer udvikler sig, skal vores tilgang til retfærdighed også gøre det. Udviklere, politiske beslutningstagere og interessenter har brug for løbende uddannelse, årvågenhed og tilpasning for at sikre, at AI-systemer forbliver retfærdige og til gavn for alle.

# CU5 | Casestudier og projekter



# Indhold

Comentado [KM17]: Must be updated in the very end (headlines and page numbers).

# 1. Introduktion

I denne kursus-enhed er der fokus på at undersøge bias i en praktisk kontekst. Dette gøres i form af et projekt, hvor der ses nærmere på et hypotetisk scenarie med en international virksomhed, der udvikler en AI-løsning til rekruttering af medarbejdere. Projektet vil blive behandlet mere detaljeret under den sidste interaktive session med de studerende. Forsøg at facilitere en livlig diskussion for at understøtte deres læring og opfordre de studerende til at udtrykke og anvende det, de har lært i de foregående sessioner.

Algoritmisk bias er fremkomsten af algoritmiske systemers uretfærdige eller diskriminerende resultater. Det sker, når en algoritme systematisk favoriserer eller forfordeler visse personer eller grupper på baggrund af specifikke karakteristika, såsom race, køn eller socioøkonomisk status. Maskinlæringsalgoritmer arbejder med forskellige datasæt, f.eks. træningsdata, hvorfra de lærer modeller, der kan anvendes på andre personer eller grupper til at komme med forudsigelser. Men disse algoritmer kan behandle lignende personer eller elementer forskelligt, hvilket fører til gentagelse eller forøgelse af menneskelige bias, hvilket især kan påvirke beskyttede grupper (Friedman & Nissenbaum, 1996; Lange & Duarte, 2018; Lee et al., 2019). At forstå årsagerne til og konsekvenserne af algoritmisk bias er afgørende for at udvikle etiske og troværdige AI-løsninger.



## Konsekvenser af algoritmisk bias

Algoritmisk bias kan optræde i forskellige former og påvirke forskellige grupper på forskellige måder. De eksempler, som Lee, Resnick og Barton (2019) har samlet, illustrerer, hvordan bias kan stamme fra flere kilder og påvirke grupper på utilsigtede eller bevidste måder.

- Ordassociationer kan afspejle bias. Forskere fra Princeton University fandt ud af, at europæiske navne blev opfattet mere positivt end afroamerikanske, og ord som "kvinde" og "pige" blev knyttet mere til kunst end til videnskab og matematik (Lee et al., 2019). Som vist i det første eksempel i dette kursus er associationer som "sygeplejerske" knyttet til kvinder og "ingeniør" er knyttet til mænd, hvilket ofte er et resultat af biased data, der er indlejret i søgealgoritmer, og som afspejler samfundsmæssige stereotyper. Dette opretholder bias inden for AI og maskinlæring på grund af afhængigheden af eksisterende online-indhold, hvilket udgør en betydelig og kontinuerlig udfordring inden for AI og maskinlæring.
- Vedrørende online-rekrutteringsværktøjer stoppede Amazon brugen af en rekrutteringsalgoritme på grund af kønsdiskriminering. AI-softwaren diskvalificerede CV'er, der indeholdt ordet "kvinder", og nedgraderede ansøgere fra kvindelige universiteter.

- Onlineannoncer udviser også bias. Harvard-forsker Latanya Sweeney opdagede, at søgninger på afroamerikanske navne havde større sandsynlighed for at resultere i annoncer, som indeholdt registre over arrestationer m.m. sammenlignet med navne på hvide personer.
- Teknologi til ansigtsgenkendelse kan også være biased. MIT-forsker Joy Buolamwini fandt ud af, at kommercielle systemer ofte ikke kan genkende mørkere hudfarver, da de hovedsageligt er trænet på lysere og mandlige ansigter.
- Selv algoritmer brugt i strafferet kan være biased. COMPAS-algoritmen, der har været brugt af dommere til kautionsafgørelser, viste sig at være biased over for afroamerikanere, som ProPublica har omtalt.

## 2. Forskellige typer af bias

Ifølge Friedman og Nissenbaum (1996) kan algoritmiske bias kategoriseres i tre hovedtyper:

- For det første er der **eksisterende bias**, som er de fordomme eller bias, der allerede findes i et system, hvad enten det er bevidst eller ubevidst, før det skabes. Disse bias kan komme ind i et system gennem en bevidst arbejdsindsats eller ubevidste handlinger, nogle gange endda med gode intentioner.
- For det andet opstår **teknisk bias** på grund af tekniske begrænsninger eller overvejelser under designprocessen. Forskellige forhold i designet kan bidrage til teknisk bias.
- For det tredje **opstår bias** i forbindelse med brug i den virkelige verden med virkelige brugere. Denne type bias opstår typisk, efter at designprocessen er afsluttet, ofte på grund af ændringer i samfundets viden, befolkning eller kulturelle værdier (Friedman & Nissenbaum, 1996).



### Eksempler på bias i datasæt til maskinlæring

Maskinlæringsmodeller er stærkt afhængige af træningsdata, men den menneskelige involvering i udvælgelsen og forberedelsen af disse data kan resultere i forskellige typer af bias. Et udbredt eksempel er **bias i rapporteringen**, hvor hyppigheden af forekomster i datasættet ikke afspejler forekomsten i den virkelige verden. Denne type bias kan føre til, at modellerne har svært ved at håndtere almindelige situationer, fordi der lægges for stor vægt på at dokumentere usædvanlige omstændigheder (Google for Developers, 2022).

En anden type er **gruppetilskrivningsbias**, som viser sig ved, at man favoriserer sin egen gruppe eller sætter andre i bås baseret på bestemte gruppekarakteristika. Implicit bias, der stammer fra éns personlige erfaringer og mentale modeller, komplicerer tingene yderligere og fører nogle gange til

**bekræftelsesbias** eller **forskerbias** under træningen af en model (Google for Developers, 2022).

**Automatiseringsbias**, hvor automatiserede systemer foretrækkes frem for menneskelig vurderinger uanset deres nøjagtighed, og **udvælgelsesbias**, der opstår som følge af ikke-repræsentative dataindsamlingsmetoder, udgør også betydelige udfordringer med hensyn til at afbøde bias i maskinlæringsmodeller (Google for Developers, 2022).

## Eksempler fra den virkelige verden

Efter at have forstået den teoretiske ramme for algoritmisk bias og dens konsekvenser er det nu passende at se nærmere på eksempler fra den virkelige verden, der illustrerer konsekvenserne af algoritmisk bias. Den første case handler om TikToks skadelige algoritme, mens den anden undersøger etiske bekymringer omkring ansigtsgenkendelsesteknologi.

## TikToks skadelige algoritme

I en afslørende undersøgelse foretaget af Finlands nationale medievirksomhed Yle kom de skadelige virkninger af TikToks algoritme frem i lyset. Yle gennemførte en undersøgelse med fokus på det indhold, der blev vist til en fiktiv 13-årig pige ved navn Ella, som led af depression og problemer med sin kropsopfattelse (YLE, 2023).



I løbet af undersøgelsen navigerede Yles dataforskere på TikTok ved hjælp af Ellas profil i fem timer. I begyndelsen så Ella muntre videoer om alt fra mad og kæledyr til vittigheder. Men inden for få minutter begyndte algoritmen at vise indhold relateret til mental sundhed, herunder antidepressiv medicin. Efterhånden som hun surfede videre, blev tonen i videoerne mørkere, og TikTok viste indhold om selvmord, selvskade og spiseforstyrrelser. Ved undersøgelsens afslutning blev hele 95 % af det indhold, Ella blev præsenteret for, anset for at være skadeligt, herunder forherligelse af spiseforstyrrelser og råd om at begrænse sit fødeindtag (YLE, 2023).

Socialpsykolog Suvi Uski fremhævede det bekymrende aspekt af algoritmens funktioner og understregede dens tendens til at prioritere brugerengagement frem for den potentielle skade, som det viste indhold forårsager. Denne afsløring understreger de betydelige etiske implikationer, der er forbundet med TikToks algoritmiske anbefalinger (YLE, 2023).

Det, der begyndte som uskyldig browser-aktivitet, eskalerede hurtigt til en bekymrende spiral af skadeligt indhold, der afspejlede Ellas sårbare tilstand. Algoritmens hurtige identifikation af Ellas interesse for depressiv og usund adfærd er alarmerende i sig selv, men endnu mere alarmerende er dens rolle i at forstærke disse skadelige tanker i stedet for at tilbyde en modvægt. I stedet for at

omdirigere Ella til positive og støttende ressource-sites forstærkede algoritmen hendes negative tendenser og forværrede hendes psykiske problemer. Det rejser kritiske spørgsmål om det etiske ansvar, som algoritmedesignere og platformsoperatører har for at sikre brugernes ve og vel. Sagen om Ella understreger det presserende behov for større gennemsigtighed, ansvarlighed og etiske overvejelser i udviklingen og implementeringen af algoritmiske systemer, især dem, der har betydelig indflydelse på brugernes adfærd og mentale sundhed.

### Uetisk indsamling af ansigtsgenkendelsesdata

Ansigtsgenkendelsesteknologi er blevet mere og mere udbredt i forskellige sektorer, fra sundhedssektoren til retshåndhævelse (Javaid, 2024). Men den udbredte anvendelse har også givet anledning til betydelige etiske overvejelser, især hvad angår dataindsamling og dens indvirkning på privatlivets fred og borgerrettigheder. Et bemærkelsesværdigt eksempel på uetisk indsamling af ansigtsgenkendelsesdata dukkede op i en rapport fra Washington Post den 7. juli 2019. Rapporten afslørede, at føderale organer som FBI og ICE havde fået adgang til kørekortdatabaser til ansigtsgenkendelsesformål og scannet millioner af amerikaneres fotos uden deres samtykke eller viden (Harwell, 2019a).

Denne foruroligende afsløring understreger det bredere spørgsmål om ansigtsgenkendelsesteknologiens potentiale for bias og diskrimination. En efterfølgende føderal undersøgelse, som også blev bragt af The Washington Post den 19. december, bekræftede forekomsten af racemæssig bias i ansigtsgenkendelsessystemer. Undersøgelsen viste, at asiatiske og afroamerikanske personer havde op til 100 gange større risiko for at blive fejlidentificeret end hvide mænd, afhængigt af den anvendte algoritme og søgetype (Harwell, 2019b).



## 3. Værktøjer til at vurdere etikken i en AI-løsning

I udviklingen af AI-systemer, især inden for områder, der er følsomme over for bias såsom rekruttering, er det vigtigt at anvende forskellige værktøjer, der er designet til at identificere og vurdere potentielle bias. Disse **vurderingsværktøjer** er afgørende for at sikre, at AI-applikationer fungerer retfærdigt og gennemsigtigt. Ved at bruge sådanne værktøjer kan udviklere bedre forstå, hvordan bias kan påvirke systemets resultater, og tage proaktive skridt til at afbøde disse effekter. Denne tilgang forbedrer ikke kun troværdigheden og retfærdigheden af AI-systemer, men hjælper også med at opbygge tillid hos brugerne ved at udvise en forpligtelse til etisk AI-praksis.

**Designværktøjer** er afgørende for varetagelse af etiske hensyn i AI-udvikling, fordi de fremmer en holistisk forståelse af forskellige interessenters perspektiver.

Værktøjerne gør det muligt for udviklere at tage indlevende stilling til AI-teknologiernes forskellige indvirkninger på forskellige brugergrupper. Ved at kortlægge disse gruppers interaktioner, adfærd og behov hjælper disse værktøjer med at identificere og afbøde potentielle bias i AI-systemer. Det brede engagement sikrer ikke kun, at AI-løsninger er mere inkluderende, men tilpasser også udviklingen til etiske standarder ved proaktivt at tage fat på problemer, der kan føre til uretfærdige eller diskriminerende resultater. I denne projektcase præsenteres forskellige værktøjer gennem hele designprocessen under hver relevant fase.

Der findes et væld af vurderingsværktøjer til dette formål, men i dette kursus fokuseres kun på at præsentere tre af dem. De designværktøjer, der blev brugt i dette projekt, er beskrevet i projektbeskrivelsen senere i denne håndbog i afsnittet "Identifikation af potentielle bias".

**UNESCO's Ethical Impact Assessment-værktøj**, der blev udviklet i 2023, tilbyder en struktureret tilgang til at vurdere de etiske konsekvenser af AI-systemer. Det består af en række forskellige spørgsmål, der omfatter både åbne spørgsmål og multiple choice-spørgsmål, og som er designet til at guide brugerne af værktøjet gennem en omfattende vurderingsproces. Værktøjet lægger vægt på etiske overvejelser helt fra starten og begynder med afgrænsende spørgsmål, før det dykker dybere ned i UNESCO's principper. Ved at tage fat på vigtige aspekter såsom påvirkninger af interessenter, team-ansvar, sikkerhed og problemstillinger vedrørende bæredygtighed, privatlivets fred, gennemsigtighed, forklarlighed og ansvarlighed giver det mulighed for en grundig undersøgelse af de etiske dimensioner, der er forbundet med udrulning af AI (UNESCO, 2023).



Delen vedrørende retfærdighed, ikke-diskrimination og mangfoldighed i UNESCO's Ethical Impact Assessment-værktøj fokuserer på at identificere algoritmisk bias, især i de data, der bruges af AI-systemer. Det undersøger grundigt de involverede roller og ansvarsområder for at sikre, at alle iboende bias håndteres (UNESCO, 2023).

Et andet værktøj er **'The Assessment List for Trustworthy AI (ALTAI)** udviklet af Ekspertgruppen på højt niveau om kunstig intelligens (High-level expert group on artificial intelligence, AI HLEG), der er oprettet af Europa-Kommissionen. Dette værktøj giver en struktureret ramme for evaluering af AI-systemers troværdighed. Værktøjet skitserer syv specifikke krav til troværdig AI med detaljerede forklaringer, der er tilgængelige på et websted, der er oprettet i denne forbindelse. Derudover indeholder det vejledning til at gennemføre evalueringen og identificerer nøgleinteressenter, der skal involveres i processen, hvilket sikrer en omfattende evaluering af AI-systemers troværdighed (ALTAI, 2020).

I ALTAI-værktøjets afsnit om diversitet, ikke-diskrimination og retfærdighed fokuseres på at undgå unfair bias. Spørgsmålene dér går på vurdering af algoritmisk bias, evaluering af inputdata og algoritmedesign for at identificere eventuel bias, samtidig med at inklusion, tilgængelighed og interessentdeltagelse fremmes for at sikre retfærdighed og gennemsigtighed i AI-systemer (ALTAI, 2020).

Det tredje værktøj, **'Mario Sosa Hidalgo's Design of an Ethical Toolkit for the Development of AI Applications'**, tilbyder en innovativ, konkret og visuel tilgang til at integrere etik i udviklingsprocessen for AI-applikationer. Værktøjet adresserer bias og råder udviklere til at implementere strategier, der minimerer uretfærdighed og sikrer, at AI-systemet er så upartisk som muligt (Hidalgo, 2019).

## 4. Projekt: En AI-løsning til rekruttering af medarbejdere

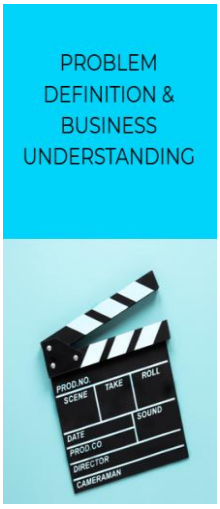
Dette afsnit fokuserer på at håndtere algoritmisk bias i AI-systemer. Det præsenterer et projekt, der involverer et hypotetisk scenarie, hvor en international virksomhed udvikler en AI-løsning til rekruttering af medarbejdere.

For at kunne gennemgå denne proces, er små fortællinger indsat i tekstbokse med blå kant, og disse fungerer som vigtige input til projektet. Derudover giver dette afsnit omfattende oplysninger for at hjælpe med at forstå konteksten og udviklingen af AI-applikationen til rekrutteringsformål.

Målet er at engagere de studerende i at genkende og afbøde bias, der kan opstå under skabelsen og implementeringen af AI-teknologier.

**Comentado [KM18]:** Can student helpers translate all blue text boxes into Danish?
Suggested translation:
Problemformulering og forretningsforståelse

133




**Baggrund**

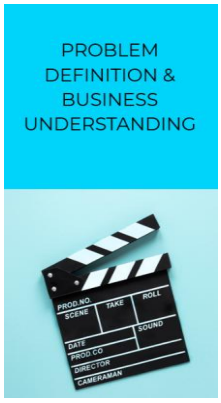
En global industrivirksomhed oplever ineffektivitet og usikkerhedsmomenter i sin rekrutteringsproces, hvilket lægger et betydeligt pres på HR-afdelingens tid og ressourcer.

I forventningen om fremtidig vækst ser virksomheden også et presserende behov for at ansætte en række specialister, herunder ingeniører, marketingfolk, finansfolk, kodere og projektledere.

Den hårde konkurrence om kompetente medarbejdere og de høje omkostninger, der er forbundet med fejlansættelser, understreger vigtigheden af en vellykket rekruttering. Situationen øger presset på HR-teamet, hvis opgave det er at sikre, at rekrutteringsprocessen bliver en succes og forløber gnidningsløst.

Kunstig intelligens (AI) bliver i stigende grad integreret i rekrutteringsprocesser og giver en række muligheder for at forbedre ansættelsespraksis. Mens AI-anvendelse i rekruttering giver betydelige fordele, giver de også visse

udfordringer. Det er afgørende for organisationer at forstå og nøje overveje flere aspekter i denne situation. Ved at forstå de potentielle fordele såvel som ulemperne kan virksomheder træffe informerede beslutninger om, hvordan de bedst implementerer AI-teknologier. Denne afbalancerede tilgang sikrer, at man udnytter kraften i AI til at strømline og forbedre rekrutteringen, samtidig med at man tager tilstrækkeligt hensyn til konsekvenser og risici.




**Forståelse af problemet**

Virksomheden begynder med at lave baggrundsundersøgelser for at udvikle en AI-løsning og danner derfor et internt team, der skal identificere de vigtigste udfordringer i den eksisterende rekrutteringsproces. Teamet gennemfører interviews med en række interessenter, herunder HR-afdelingen, nyansatte medarbejdere og ledere, der for nylig har ansat nye teammedlemmer. Teamet indsamler også data om de nuværende rekrutteringsomkostninger og analyserer succesraten for disse tiltag.

Det er afgørende for virksomheden at forstå potentielle medarbejderes opfattelse af brugen af AI i rekrutteringen. Den ønsker at undgå skader på sit omdømme eller brand. For at opnå dette er virksomheden forpligtet til at sikre, at indførelsen af kunstig intelligens ikke kun får bred accept, men også overholder etiske og juridiske standarder.

**Efter diskussioner og undersøgelser bliver fordele og ulemper ved en AI-rekrutteringsløsning som følger:**



## Fordele ved brug af AI i rekrutteringsprocesser

Kunstig intelligens revolutionerer rekrutteringsprocessen ved at automatisere rutineopgaver, så de rekrutteringsansvarlige kan koncentrere sig om mere strategiske aspekter af deres arbejde. Denne automatisering forbedrer ikke kun den samlede effektivitet i håndteringen af ansøgninger, men reducerer også den tid, der bruges på rekrutteringsprocesser, betydeligt (Albaroudi et al., 2024; www.recruiter.com, 2023).

Desuden gør AI's evne til at analysere omfattende datasæt det muligt at identificere de bedst egnede ansøgere og dermed forbedre kvaliteten af ansættelserne. Denne analytiske evne kan potentielt reducere personaleomsætningen ved at sikre et godt match mellem jobbet og kandidaten. Den hurtige feedback og de opdateringer, som AI muliggør, forbedrer ikke kun kommunikationen, men holder også ansøgerne velinformerede gennem hele

udvælgelsesprocessen. Det giver de rekrutteringsansvarlige mere tid til at engagere sig i udvalgte kandidater, hvilket yderligere forbedrer ansøgernes oplevelse (Arivu Recruitment and Consulting, 2023; Sheard, 2022; www.recruiter.com, 2023).

Implementeringen af AI i rekrutteringen fører også til omkostningsreduktion ved at minimere afhængigheden af manuel arbejdskraft (www.recruiter.com, 2023). Derudover forsøger AI'en at mindske menneskelige bias og dermed øge objektiviteten, ensartetheden og retfærdigheden i rekrutteringsprocessen. Dette teknologiske fremskridt kommer især de grupper til gode, der måtte opleve barrierer for økonomisk deltagelse, da det potentielt forbedrer deres ansættelsesmuligheder (Bursell & Roumbanis, 2024; Köchling & Wehner, 2020; Sheard, 2022).

AI's evne til skarpe analyser strækker sig til screening af sociale medier, hvor den vurderer kandidaternes profiler med hensyn til, hvordan de passer til jobfunktionen. Den screener også for upassende indhold og sikrer, at rekrutteringen er i overensstemmelse med organisationens værdier. Desuden hjælper AI-værktøjer med at overvinde sprogbarrierer i rekrutteringen, hvilket giver adgang til en bredere og mere mangfoldig talentpulje på globalt plan (Albaroudi et al., 2024).

Endelig forventes AI at forbedre mangfoldigheden på arbejdspladsen ved at fjerne bias fra ansættelsesprocessen (Sheard, 2022). Denne række af fordele understreger den transformerende effekt af AI i rekrutteringsprocesser og giver mulighed for en mere effektiv, retfærdig og inkluderende ansættelsesproces.



## Betænkeligheder ved brug af AI i rekrutteringsprocesser

Indførelsen af kunstig intelligens i rekrutteringsprocessen giver mange effektivitetsgevinster, men medfører også en udfordring i form af begrænset menneskelig kontakt og upersonlig interaktion. Denne reduktion i det personlige engagement kan få rekrutteringsoplevelsen til at føles mekanisk, hvilket kan få kandidater til at føle sig fremmedgjorte og forringe kvaliteten af rekrutteringsoplevelsen (www.recruiter.com, 2023).

Desuden udgør risikoen for bias og diskrimination et betydeligt opmærksomhedspunkt. Hvis AI ikke er ordentligt designet og trænet, eller hvis den er baseret på forudindtagede træningsdata, kan den utilsigtet indføre bias i rekrutteringsprocessen. Det kan have en negativ indvirkning på mangfoldighed, inklusion og retfærdighed i ansættelsespraksis (Albassam, 2023; www.recruiter.com, 2023).

Problemer med databeskyttelse og sikkerhed er også altafgørende i betragtning af de store mængder persondata, der indsamles og behandles af AI-systemer. Dette rejser spørgsmål om privatlivets fred, databeskyttelse og risikoen for cyberangreb, hvilket understreger behovet for robuste

sikkerhedsforanstaltninger og gennemsigtighed i dataanvendelsen (Albassam, 2023; www.recruiter.com, 2023).

Implementeringen af kunstig intelligens i rekrutteringen er ikke uden omkostninger, og det er ressourcekrævende. Især for små virksomheder kan den nødvendige økonomiske og ressourcemæssige investering være betydelig, hvilket udgør en adgangsbarriere, som kan begrænse indførelsen af AI i deres rekrutteringsprocesser (www.recruiter.com, 2023).

Der er også opmærksomhedspunkter med hensyn til manglen på menneskelig dømmekraft og potentielle problemer med nøjagtigheden. Hvis man udelukkende stoler på AI, uden at rigtige mennesker foretager tilsyn af, kan det resultere i, at man overser kvalificerede kandidater på grund af ufuldstændige data eller færdigheder, der ikke er beskrevet tilstrækkeligt fyldestgørende eller korrekt. Desuden kan AI's nøjagtighed blive kompromitteret af ansøgere, der manipulerer deres data for at virke mere kvalificerede (Albassam, 2023; Arivu Recruitment and Consulting, 2023).

De skiftende juridiske rammer omkring brugen af kunstig intelligens i rekruttering kræver, at virksomhederne tilpasser sig og overholder dem for at sikre retfærdighed i hele processen. Denne fortsatte udvikling understreger behovet for konstant årvågenhed og tilpasning med hensyn til brugen af AI-værktøjer i rekrutteringsprocessen (Fernandes, 2021; www.recruiter.com, 2023).



Begrænsninger i AI's evne til at vurdere, om ansøgere passer ind i virksomhedskulturen, eller hvordan deres samarbejdsevner er, kan resultere i, at man går glip af muligheder for at ansætte kandidater, der kunne vise sig at være ganske udmærkede, hvis de fik chancen. Albassam (2023) understreger vigtigheden af at overveje disse faktorer, som ofte er afgørende for, hvordan en ny medarbejder vil præstere, men som er vanskelige for AI-systemer at kvantificere nøjagtigt.

Markedsdominans og mangel på gennemsigtighed i AI-ansættelsesværktøjer, især dem, der er udviklet af private virksomheder i USA, komplicerer indsatsen for at mindske bias og ekstern kontrol. Disse systemers proprietære eller privatejede karakter, såsom intellektuel ejendom, gør det udfordrende at vurdere og forbedre deres retfærdighed og effektivitet (Sheard, 2022).

Endelig er forværringen af ulighed på grund af automatiserede ansættelsesplatforme en voksende bekymring. Disse systemer kan forstærke eksisterende sociale uligheder gennem biased algoritmer og gennem den vigtige rolle, som netværk og uformelle strukturer spiller i forbindelse med jobsøgning. Denne bekymring peger på behovet for omhyggelige overvejelser og handlinger for at fjerne bias fra AI-drevne ansættelsesplatforme (Harvard John A. Paulson School of Engineering and Applied Sciences, 2023).

Hvert af disse punkter understreger kompleksiteten og de mange aspekter, der er, ved at inkorporere AI i rekrutteringsprocesser, og punkterne tydeliggør

behovet for en afbalanceret tilgang, der inddrager effektivitet, retfærdighed og inklusivitet.

## 5. Identificering af potentielle bias


**Forståelse af systemets brugere**

En af de interviewede personer er *Lead Recruiter* Maria, som har flere års erfaring i virksomheden. Hun har med egne øjne set den stadige ophobning af rekrutteringsopgaver og det voksende pres, det lægger på HR-afdelingen, og som intensiveres, når man ansætter internationale talenter.

Maria er åben og interesseret i mulighederne for at bruge kunstig intelligens til at hjælpe med rekrutteringsopgaver, selv om hun ikke har et indgående kendskab til teknologien i kunstig intelligens.



### Værktøj: Persona

Personaer bruges i designfasen til at skabe detaljerede profiler, der repræsenterer specifikke typer af brugere eller grupper. Disse profiler indfanger væsentlige karakteristika, uden man griber til egentlige stereotyper. Ved at inkludere realistiske detaljer såsom adfærd, behov og demografiske oplysninger bliver personaer relaterbare og nyttige til at forstå forskellige brugergrupper. De hjælper design- og udviklingsteams med at opbygge empati, tilpasse deres arbejdsindsats og udvikle løsninger, der er skræddersyet til at opfylde specifikke behov. Personaer er særligt anvendelige, fordi de hjælper med at nedbryde traditionelle marketingsegmenter, hvilket fører til et mere inkluderende og effektivt design (Kumar, 2013; Stickdorn et al., 2018a, 2018b).

### HR-medarbejder Marias forståelses-bias

Når man vurderer personaen, er det vigtigt at tage i betragtning, at hendes personlige synspunkter kan påvirke de anbefalinger og holdninger, hun udtrykker om brugen af værktøjet. Der er en løbende debat om, hvorvidt AI udgør en udfordring eller risiko for traditionelle rekrutteringsroller. HR-

medarbejdere kan opfatte AI-rekrutteringssystemer som en trussel mod deres job, hvilket kan resultere i en modvilje mod at anvende AI-værktøjer til rekruttering (Chen, 2023). Andre personlige fordomme baseret på hendes persona kunne f.eks. være følgende:

- Åbenhed eller skepsis, relateret til alder, over for ny teknologi og dens indvirkning på traditionelle jobfunktioner.
- Indflydelse fra at være i en teknologisk fremadstormende by som Helsinki, muligvis med en tendens til innovative løsninger.
- Hendes guatemalanske baggrund kan påvirke hendes opfattelse af AI's indvirkning på mangfoldighed og inklusion i rekrutteringen.
- Hendes uddannelse i HR-ledelse kan have fremmet en tro på vigtigheden af menneskelig dømmekraft, hvilket muligvis forårsager modvilje mod at stole på AI til vigtige rekrutteringsopgaver.
- Når hun bor alene, kan hendes personlige erfaringer få hende til at værdsætte menneskelig interaktion mere, hvilket fører til skepsis over for AI's upersonlige væsen.
- I hendes målsætninger på arbejdspladsen lægger hun vægt på effektivitet og personligt engagement i kandidaterne, hvilket kan komme i konflikt med hendes syn på AI-rekruttering.
- Frustrationer over store arbejdsbyrder og langsomme processer kan gøre hende åben over for AI's effektivitet, men frygten for, at det hele bliver upersonligt, og for at miste kontakten med kandidaterne kan få hende til at tøve.
- Hobbyer som svømning, klatring og læsning indikerer en afbalanceret, mangefacetteret tilgang til livet, hvilket i hendes perspektiv også kan betyde et ønske om en nuanceret tilgang til rekruttering.
- Bekymringer om, at AI ændrer HR-funktioner, kan få hende til at sætte spørgsmålstegn ved, hvordan det passer til hendes kompetencer og det fremtidige behov for hendes jobfunktion.



## Rekrutteringsrejse


**Forstå rekrutteringsrejsen**

For at få en omfattende forståelse af rekrutteringsprocessen udvikler virksomheden en rekrutteringsrejse, hvor Maria og rekrutteringsteamet indgår. I samarbejde med linjeledere navigerer de gennem hvert trin på rejsen for at identificere væsentlige frustrationer og tidskrævende trin. Derudover undersøger de den potentielle integration af AI på hvert trin for at forbedre rekrutteringsrejsens effektivitet.

### Værktøj: Kort over rejsen

Et kort over rejsen er et visuelt værktøj, der bruges i designet til at skildre den trinvise oplevelse, en bruger har med en service. Det repræsenterer brugerens perspektiv og beskriver, hvad der sker på hvert interaktionsstadie, de involverede touchpoints og eventuelle udfordringer eller forhindringer, de måtte stå over for. Derudover indeholder et kort over rejsen ofte 'lag', der viser brugerens følelsesmæssige reaktioner, hvorved positive eller negative oplevelser i løbet af deres rejse tydeliggøres. Disse kort kan bruges til at illustrere både aktuelle oplevelser (*current-state journey maps*) og forventede fremtidige interaktioner (*future-state journey maps*). De varierer i omfang, fra brede, generelle oversigter på højt niveau over hele oplevelsen til meget detaljerede kort med fokus på specifikke øjeblikke (Kumar, 2013; Stickdorn et al., 2018a).

I dette projekt introduceres kortet over rejsen først som en beskrivelse af den nuværende tilstand, efterfulgt af en fremskrivning af den fremtidige tilstand, der skitserer de muligheder, som AI-værktøjer kan give i rekrutteringsprocesser. Brugen af AI i forskellige rekrutteringsfaser behandles mere detaljeret i afsnittet "Præsentation af AI-løsninger: Håndtering af rekrutteringsprocessens faser og deres tilhørende bias".

**Comentado [KM19]:** Nidhi: If these illustrations should be translated, can you facilitate that?


Comentado [KM20]: Student helpers: Please translate the blue box 'Designfase' since I am not able to edit the text.


**At forstå personlige bias**

Efter en klar problemformulering har virksomheden valgt at gå videre med en AI-drevet løsning. Du er blevet udnævnt til ledende AI-udvikler med ansvar for at føre tilsyn med implementeringen af den valgte AI-teknologi.

Virksomheden forstår vigtigheden af at samle en mangfoldig gruppe med flere interessenter for at sikre, at en række forskellige perspektiver indgår, hvorved projektet gerne skulle blive en succes.

I et forsøg på at håndtere potentielle, ubevidste bias i teamet kræver virksomheden, at alle gruppemedlemmer deltager i en øvelse, baseret på Alan Turing, som kaldes "Positionality Matrix".

## Værktøj: Alan Turings "positionality matrix"



At reflektere over "positionalitet" – dvs. at forstå ens sociale og kulturelle baggrund – er afgørende for AI-udvikling og etisk beslutningstagning. Alan Turing Institute (2022) understreger vigtigheden af "positionalitetsmatrixen", et værktøj til at vurdere, hvordan individuelle træk, såsom alder, race og uddannelse, påvirker forskning og innovation. Denne selvbevidsthed er nøglen til at forhindre bias og magtubalance i AI-systemer. Ved at overveje disse personlige faktorer kan udviklere sikre, at AI-løsninger er retfærdige og respekterer de forskellige samfund, de servicerer, og at de dermed bidrager positivt uden at opretholde uretfærdig praksis.

Alan Turings *positionality matrix* er en ramme til refleksion, der afslører indflydelsen fra individuelle socioøkonomiske baggrunde, kulturelle kontekster og livserfaringer på dømmekraft og beslutningstagning i et team. Den fungerer som et værktøj til at skelne mellem og håndtere indflydelsen af forskellige personlige karakteristika, såsom alder og etnicitet, på forsknings- og udviklingsperspektiver. Denne matrix er medvirkende til at afdække ubevidste fordomme og navigere i magtdynamikker og derved fremme en forpligtelse til at sikre mangfoldighed og retfærdighed i AI-udvikling (Alan Turing Institute, 2022).

**At tage hensyn til interessenter**


Ud over at skulle tage hensyn til den interne gruppe af interessenter, der deltager i projektet, er virksomheden også nødt til at tage hensyn til andre relaterede interessenter.

Derfor udarbejdede virksomheden et kort over interessenter, inspireret af G.J. Millers definition af kortlægning af interessenter i en AI-udviklingsproces, og dette kort viser udviklings- og brugerinteressenterne samt de eksterne interessenter.

På baggrund af dette kort skal projektlederen inddele interessenterne i grupper for at vide, hvordan de skal holdes informeret, og for at se, hvilken indflydelse de har på projektet.

### Værktøj: Kort over interessenter

Et kort over interessenter er et visuelt værktøj, der bruges til at identificere og vise rollerne og relationerne mellem alle interessenter, der er involveret i et projekt. Det kan variere fra en simpel kvadrant, der viser niveauer af indflydelse og engagement, til en detaljeret matrix, der beskriver interaktioner mellem interessenter. Kortet over interessenter hjælper med at tydeliggøre interessenternes dynamik og er afgørende for at forstå og styre netværksrelationer (Giordano et al., 2018; SDT, n.d.).

### Værktøj: Matrix til analyse af interessenter

Interessentanalysematrixen er et visuelt værktøj, der bruges til at kategorisere og håndtere interessenter i et projekt baseret på deres indflydelses- og interesseniveau. Ved at indplacere interessenter i et kvadrantnet hjælper denne matrix med at prioritere strategier med hensyn til engagementet. Kvadranten øverst til venstre repræsenterer interessenter med stor indflydelse, men lav interesse, hvilket kræver en indsats for at holde dem tilfredse uden at overvælde dem. Kvadranten øverst til højre omfatter meget indflydelsesrige og interesserede interessenter, som kræver tæt styring. Den nederste højre kvadrant, som viser interessenter med stor interesse, men mindre indflydelse, bør få regelmæssige opdateringer for at holde dem informeret. Endelig kræver den nederste venstre kvadrant, som er for dem med minimal interesse og indflydelse, minimalt engagement; stort set bare at man holder lidt øje med dem (Hoory & Botorff, 2022; Wallbridge, 2023).


**Bias i rekrutteringsfasen – workshop**

Når interessenterne er klarlagt, beslutter virksomheden at afholde en workshop om etik, hvor man gennemgår alle mulige bias i de forskellige ansættelsesfaser.

Interessenter fra området vedr. "Samarbejde" (forrige slide) inviteres til at deltage.

Hvilken slags Post-it-sedler ville du have udfyldt?

### Værktøj: Workshop

En workshop er et kollaborativt designværktøj, hvor en gruppe deltagere indgår i interaktive aktiviteter for at løse et specifikt problem eller arbejde på et projekt. Workshops faciliteres typisk af en leder og kan variere i varighed fra et par timer til flere dage, afhængigt af målene og opgavernes kompleksitet. Denne metode er effektiv til at generere ideer, løse problemer og fremme teamwork og innovation i en gruppe (Wirtz, 2022).



## Præsentation af AI-løsninger: Håndtering af rekrutteringsprocessens faser og deres tilhørende bias

Algoritmer og maskinlæringsteknologier bruges i stigende grad – helt fra de indledende faser af rekrutteringsprocessen. Disse forudsigende tilgange hjælper med at foregribe forretningsudviklingen ved at analysere eksterne faktorer som efterspørgsel, konkurrenters nye træk og skiftende forbrugerbehov. HR-metrikker spiller en afgørende rolle i udarbejdelsen af disse prognoser. De omfatter en analyse af tidligere tendenser i arbejdsstyrken, demografiske ændringer og nye krav til kompetencer. Derudover er interne vurderinger, som f.eks. vurdering af arbejdsbyrden, ledelsesobservationer, medarbejderfeedback samt arbejdsstyrkens demografi, afgørende for at forstå den aktuelle situation og den løbende udvikling af medarbejderstaben. Målinger som f.eks. omsætningshastighed mht. medarbejdere, tiden der går med at besætte stillinger og omkostninger pr. ansættelse er afgørende for at måle en organisations rekrutteringseffektivitet og identificere områder, hvor processen kan forbedres (www.recruiter.com, 2023).

Bias i forudsigelige ansættelser (dvs. når man bruger AI til at screene jobansøgninger og forudsige, hvem der er de bedste kandidater) kan opstå fra forskellige kilder. Disse kilder kan f.eks. være præstationsniveauet for nuværende medarbejdere i specifikke roller, hvilket kan føre til forankringsbias. Hvis f.eks. en medarbejder, der er ualmindelig dygtig, skal udskiftes, vil analyserne måske foreslå én ny medarbejder. Men den afgående medarbejders høje præstationsniveau kan faktisk nødvendiggøre to nye medarbejdere. Omvendt kan en medarbejder, som præsterer på et lavt niveau, fejlagtigt antyde, at der er behov for yderligere ansættelser for at kompensere for den manglende produktivitet (Chen, 2023, s. 145). Derudover kan modstand mod forandring blandt medarbejdere føre til en status quo-bias, hvor man er tilbageholdende med at tilpasse organisationsstrukturer, selvom der klart er et behov for dette i virksomheden (Sukernek, 2024).

Desuden kan forudsigelige analyser nogle gange overse vigtige faktorer på grund af et for snævert fokus. Hvis man f.eks. ikke tager højde for mere generelle trends i en branche eller demografiske ændringer, som f.eks. en aldrende befolkning, kan en virksomhed være helt uforberedt på fremtidige krav til arbejdsstyrken.

Ligesom andre trin i rekrutteringsprocessen er denne fase også påvirkelig over for feedback loop-bias. Denne type bias kan opstå, når AI-systemer, der bruges til rekruttering, trænes på historiske data og utilsigtet forstærker systemernes egne fordomsfulde beslutninger. Dette kan ske i dynamiske miljøer, hvor AI'ens beslutninger påvirker de data, den vil lære af i fremtiden. Hvis en AI, der allerede er biased af tidligere data, fortsætter med at træffe skæve ansættelsesvalg, føres disse beslutninger tilbage til dens læringsdatasæt, hvilket viderefører og endda forstærker de oprindelige bias (Vivek, 2023, s. 0107).



## Bias i jobbeskrivelsen

Kunstig intelligens kan gøre det lettere at lave præcise jobbeskrivelser. Ved at tilpasse sproget i annoncer og overvåge effekten af disse ændringer på mængden og kvaliteten af ansøgninger kan organisationer forbedre effektiviteten af deres promoveringsindsats. Desuden kan AI bestemme, hvilke aspekter af en virksomhed, herunder dens kultur og resultater, der skal fremhæves over for potentielle kandidater for at få den bedste respons (Chen, 2023, s. 140).

Jobbeskrivelser spiller en afgørende rolle, når en virksomheds behov skal formidles til potentielle medarbejdere. Men bias i disse beskrivelser kan forhindre kvalificerede personer i at søge. Sådanne bias drejer sig ofte om alder, køn eller uddannelsesmæssige kvalificationer (Ongig, 2024).

Interessant nok antyder artikler, at AI har potentiale til at reducere nogle af disse problemer ved at gennemgå jobbeskrivelser og identificere sprogbrug, der kan bidrage til bias. Selvom AI-genererede jobbeskrivelser måske udviser mindre bias end menneskeskabte, er der stadig risiko for at fastholde eksisterende fordomme,

hvis AI-værktøjerne ikke er omhyggeligt designet (CareerExperts, 2023). For eksempel kan der opstå forankringsbias fra en leders tidligere ansættelsesbeslutninger, hvor de nuværende, top-præsterende medarbejderes egenskaber og kompetencer påvirker udarbejdelsen af nye jobbeskrivelser (Chen, 2023, s. 145).

Derudover kan AI's forslag om at bruge bestemte udtryk i jobannoncer, som "ambitiøs" eller "selvsikker leder", utilsigtet favorisere eller afskrække bestemte demografiske grupper, og dermed forårsage, at bias opstår (Lawton, 2022). Hvis AI-systemer trænes ved hjælp af tidligere jobbeskrivelser, risikerer de at gentage sproglige vendinger og krav, der historisk set har tiltrukket en mindre mangfoldig pulje af ansøgere, hvilket potentielt kan fortsætte denne tendens.

## Bias i talentsøgning

Der anvendes forskellige AI-metoder i den indledende kontaktfase med potentielle kandidater. Algoritmer kan gøre det lettere at kommunikere på den helt rigtige måde på tværs af digitale platforme og sociale netværk, mens AI-bots bruges til at finde både aktive og passive jobsøgere på tværs af forskellige platforme eller endda blandt en virksomheds database med tidligere ansøgere. Derudover kan AI anvende tekstmining-strategier til at gøre jobopslag mere attraktive ved at indarbejde sætninger, der tiltrækker flest kandidater. Desuden kan et AI-system vurdere tonefaldet i en tekst og give anbefalinger til at forfine jobbeskrivelserne og gøre dem mere inkluderende (Hunkenschroer & Luetge, 2022, s. 992).



AI-drevne jobannoncer indebærer i sagens natur en risiko for indirekte diskrimination. Det skyldes, at arbejdsgivere kan specificere målgruppen for deres annoncer ved hjælp af kriterier som sprogfærdigheder, uddannelsesmæssig baggrund, erhvervserfaring eller endda alder og køn, hvilket kan forhindre visse grupper i overhovedet at se disse jobmuligheder, og hvilket igen bidrager til homogenitet på en arbejdsplads (Chen, 2023, s. 144; Sheard, 2022, s. 624) . For eksempel kunne Verizons beslutning om at målrette en Facebook-annonce mod personer i alderen 25 til 36 år, der bor i eller har besøgt den amerikanske hovedstad, og som er interesseret i finans, utilsigtet udelukke andre potentielle jobsøgere fra annoncens dækningsområde (Sheard, 2022, s. 624). Algoritmer, der er designet til at forbedre effektiviteten af jobannoncer, har en tendens til at fokusere på demografi, der historisk set interagerer mere med disse annoncer, og dermed indsnævres mangfoldigheden i ansøgerpuljen over tid (Köchling & Wehner, 2020).

Rekrutteringsbias forstærkes yderligere af rekrutteringsmedarbejdernes præferencer for specifikke annonceringsplatforme og deres tendens til at engagere sig mere med kandidater fra disse platforme. Det kan føre til manglende mangfoldighed, da forskellige platforme tiltrækker forskellige typer brugere (Lawton, 2022). Bias i rekrutteringsprocesser er ikke kun et resultat af algoritmer eller arbejdsgiverpraksis; det stammer også fra ansøgernes egen

adfærd. Hvis man bruger sociale medier eller andre onlineaktiviteter til at finde potentielle kandidater, kan det give en skævvridning i retning af dem, der er aktive online eller har knowhow til at skabe en attraktiv digital profil. Det er især tydeligt i meget konkurrenceprægede brancher som IT, hvor de travleste fagfolk måske ikke opdaterer deres LinkedIn-profiler regelmæssigt. Omvendt er der dem, der måske strategisk styrer deres onlineprofiler for at øge deres tiltrækningskraft, hvilket rejser spørgsmål om ægtheden af sådanne digitalt kuraterede personaer (Naakka, 2018, s. 46-47).

Spørgsmålet om gennemsigtighed er et væsentligt, etisk bekymringspunkt i AI-assisteret rekruttering. Algoritmernes kompleksitet kan resultere i en "black box"-effekt, hvor begrundelserne for beslutningerne forbliver uklare for ansøgerne (DigitalOcean, n.d.). Derudover er der de etiske spørgsmål vedrørende brugen af kandidaters profiler og aktiviteter på sociale medier og forskellige platforme til ansættelsesformål. Standarderne for håndtering af sådanne data varierer internationalt, og den europæiske generelle databeskyttelsesforordning (GDPR) er blandt de strengeste (Hunkenschroer & Luetge, 2022, s. 995).

## Bias i screening og udvælgelse af kandidater

AI-teknologier spiller en afgørende rolle i den indledende screening af ansøgere ved at prioritere dem, der ser ud til at være bedst egnede til jobbet. Historisk set har virksomheder været afhængige af at scanne CV'er for specifikke nøgleord. Men moderne tilgange er mere sofistikerede, herunder chatbots og værktøjer til analyse af CV'er, der søger efter semantiske sammenhænge og vurderer andre kvalifikationer. Nogle værktøjer forudsiger endda en kandidats potentielle fremtidige præstationer ved at lede efter indikatorer for vedkommendes engagement eller produktivitet på længere sigt samt fraværet af negative markører som tendens til at komme for sent eller problemer med disciplinen (Hunkenschroer & Luetge, 2022, s. 992).

I processen med at udvælge kandidater anvendes AI også til at analysere videointerviews. AI-assistenter gennemfører disse interviews, stiller en række spørgsmål og evaluerer ikke kun svarene, men analyserer også kandidatens tonefald, ansigtsudtryk og følelsesmæssige reaktioner for at nå frem til nogle personlighedstræk. Derudover bruges AI-drevne videospil til at evaluere karaktertræk såsom risikovillig adfærd, planlægningsevner, udholdenhed og motivation (Hunkenschroer & Luetge, 2022, s. 992).

Desuden analyserer AI-løsninger ansøgernes digitale fodspor, herunder aktiviteter på sociale medier og anden onlineadfærd, ved hjælp af sproglig analyse for at vurdere, hvordan de passer ind i virksomhedens kultur (Hunkenschroer & Luetge, 2022, s. 992). Ud over at undersøge sociale medier anvendes neuro-lingvistiske programmeringsteknikker (NLP) i forskellige rekrutteringsfaser, herunder CV-analyse, for at opnå en dybere viden (Albaroudi et al., 2024, s. 391).

AI har potentiale til at skabe et mere retfærdigt rekrutteringsmiljø ved at reducere ubevidst bias. Ved at fokusere på foruddefinerede evner og referencer kan AI-drevne løsninger se bort fra demografiske detaljer som alder, køn og etnicitet, som utilsigtet kan påvirke en rekrutteringsmedarbejders vurdering. En sådan tilgang har potentiale til at forbedre mangfoldigheden på arbejdspladsen, da den sikrer, at kandidater vurderes ud fra deres kvalifikationer og resultater. Desuden er visse AI-applikationer specifikt udviklet til at hjælpe virksomheder med at nå deres mål for mangfoldighed og inklusion. De opnår dette ved at undersøge rekrutteringstendenser og foreslå justeringer for at afhjælpe eventuelle uligheder (DigitalOcean, n.d.).

På trods af screeningsprocessernes potentiale til at reducere visse bias kan de utilsigtet introducere andre, hvis de ikke styres omhyggeligt. Bias kan stamme fra principperne bag modeldesign, udvælgelsen af funktioner og de data, der bruges til træningen (Hunkenschroer & Luetge, 2022, s. 994).

Når træningsdata er baseret på en pulje af eksisterende medarbejdere eller en snævert defineret gruppe, der ikke proportionelt repræsenterer forskellige befolkningsgrupper, kan det føre til utilsigtet diskrimination af underrepræsenterede grupper (Hunkenschroer & Luetge, 2022, s. 994). AI-systemer, der lærer af data, som måske allerede indeholder historiske bias, kan forstærke disse bias yderligere. Hvis et AI-system f.eks. overvejende er trænet på CV'er fra en bestemt demografisk gruppe, kan det uretfærdigt favorisere kandidater fra denne gruppe (Vivek, 2023). Et bemærkelsesværdigt tilfælde opstod i 2018, da Amazons AI-ansættelsessystem viste præference for mønstre, der var dominerende blandt mandlige ansøgere, og dermed diskriminerede kvindelige kandidater. Det skyldtes, at Amazons træningsdata primært bestod af CV'er fra top-performere, hvoraf de fleste var mænd, hvilket fik algoritmen til at forbigå egenskaber, der typisk forbindes med kvinder (Albaroudi et al., 2024, s. 386; Hunkenschroer & Luetge, 2022, s. 994).

Desuden kan der også opstå uddannelsesmæssige bias, når et algoritmisk værktøj ikke tager alle de færdigheder og egenskaber i betragtning, der er relevante for jobbet. Hvis værktøjet fokuserer for meget på tekniske kompetencer uden at tage højde for interpersonelle kompetencer, kan vigtige kompetencer som kommunikationsevner blive overset. Aggregeringsbias, hvor der foretages falske generaliseringer om hele populationer, kan også forekomme og føre til udelukkelse af enkeltpersoner baseret på antagelser om gruppekarakteristika. For eksempel kan algoritmer antage, at yngre mennesker har de allerbedste teknologiske færdigheder, og dermed overse ældre kandidater, der er lige så kvalificerede (Albaroudi et al., 2024, s. 386).

Vurderingen af en kandidats digitale fodspor som en del af screeningen kan også fastholde bias, der afspejler de forskellige opfattelser af, hvad der bør præsenteres på sociale medier, og de begrænsninger, folk står over for, når de skal opdatere deres profiler. En undersøgelse fra North Carolina State University med interviews af 61 HR-medarbejdere afslørede en tendens til at lægge vægt på personlige interesser, som f.eks. at vandre eller fejre jul, frem for faglig kunnen. Denne

tilgang kan være en fordel for visse demografiske grupper frem for andre, og i tilgangen antages det, at egenskaber som at være "energisk" eller "aktiv" i sagens natur er bedre, hvilket potentielt kan diskriminere ældre eller handicappede personer. Ved at indføre AI i denne proces risikerer man at indlejre disse fordomme i algoritmerne (Shipman, 2021).

Ved brug af software til registrering af følelser i videoer, hvor man vurderer ansøgere, skal man nøje overveje de forskellige intonationer på tværs af sprog og potentialet for systematisk at forfordele visse racer eller etniske grupper. Desuden kan nogle personer finde det udfordrende at opføre sig naturligt foran et kamera, hvilket fører til, at AI-værktøjet foretager unøjagtige vurderinger (Hunkenschroer & Luetge, 2022, s. 995).

Fortrolighed og informeret samtykke ved brug af oplysninger om ansøgere i forbindelse med rekruttering er vigtige etiske overvejelser, og reglerne varierer internationalt. Den europæiske persondataforordning (GDPR) er blandt de strengeste rammer og er designet til at beskytte EU-borgernes rettigheder ved at regulere indsamling, opbevaring og behandling af persondata og kræve informeret samtykke til enhver behandling af persondata. Ikke desto mindre opstår et dilemma, hvis ansøgere vil fravælge det af frygt for, at det kan bringe deres jobmuligheder i fare (Hunkenschroer & Luetge, 2022, s. 995).



## Bias i jobsamtaler

En af grundene til at bruge AI-baserede interviewteknikker kan være sprogbarriererne i multinationale rekrutteringsforløb, som kræver lokalt kendskab og flersprogede HR-medarbejdere. AI afhjælper dette ved at gøre det muligt at kunne interviewe personer, der taler forskellige sprog, uden at HR behøver at kende alle lokale dialekter, takket være fremskridt inden for naturlig sprogbehandling (NLP). Denne teknologi gør det muligt for virksomheder effektivt at bygge bro over sprogkløften ved internationale ansættelser uden at være afhængige af oversættere (Albaroudi et al., 2024, s. 392).

AI-baserede interviewværktøjer bruges dog oftere til at automatisere den første interviewrunde og på den måde bidrage til vurderingen af kandidater (Fritts & Cabrera, 2021, s. 792). Derfor behandles AI-baserede interviewløsninger og de mulige bias, de skaber, i kapitlet "Screening og udvælgelse af kandidater".

Vivek (2023) antyder, at på grund af den nuancerede forståelse og følelsesmæssige intelligens, som mennesker besidder, bør den endelige beslutningskompetence forblive hos 'rigtige' rekrutteringsmedarbejdere. Fritts & Cabrera (2021) nævner også, at bekymringerne omkring AI-drevne processer til identifikation, screening, interview og onboarding af kandidater kan føre til dehumanisering ved at fjerne den personlige interaktion. Dette kan være etisk problematisk, da de kunstige værdier, der er indlejret i ansættelsesalgoritmer, kan underminere forholdet mellem medarbejder og arbejdsgiver ved at fjerne de iboende menneskelige aspekter af traditionel, menneskelig interaktion.

## Bias i evalueringsproces og præsentation af jobtilbud

Når virksomheden har valgt en kandidat, skal den give vedkommende et jobtilbud. AI-algoritmer kan i den forbindelse evaluere kandidatens tidligere stillinger for at formulere et jobtilbud, som kandidaten sandsynligvis vil acceptere. Men disse værktøjer kan forværre eksisterende uligheder med hensyn til begyndelsesløn og løbende aflønning blandt forskellige køn, racer og andre demografiske grupper. Det sker, fordi disse værktøjer ofte er afhængige af historiske løndata, som kan indeholde iboende bias (Lawton, 2022).

## Bias i onboarding

Implementering af AI i onboarding kan reducere HR-medarbejdernes rutineopgaver ved at automatisere dokumenthåndtering, planlægge kommunikation og bruge chatbots til indledende feedback og forespørgsler. Et AI-værktøj kan også vurdere nye medarbejderes fremskridt og sikre overholdelse af uddannelseskrav, samtidig med at det giver analyser af onboarding-tendenser (Marr, 2023).

AI-interfaces kan skræddersy uddannelsesforløb, så de passer til individuelle færdigheder og læringsstile, hvilket sikrer, at nyansatte får den støtte, de har brug for til at få succes i jobbet. HR-afdelinger kan bruge viden fra automatiserede interaktioner og feedback til at forstå nye ansattes evner og vækstområder. Forudsigende analyser kan forudse nye medarbejderes behov og tilbyde dem ressourcer og støtte, før behovene er opstået (Marr, 2023). AI kan også bruges til at matche den nyankomne med en mentor på arbejdspladsen baseret på deres færdigheder, interesser og ekspertise (Featured, 2023).

Ligesom i tidligere faser kan bias i AI-drevet onboarding opstå som følge af brugen af træningsdata, der udspringer af historiske optegnelser, som indeholder tidligere fordomme. Desuden kan algoritmerne utilsigtet inkorporere udviklernes ubevidste fordomme i udformningen af uddannelsesforløb.

## 6. Screening og udvælgelse af kandidater: definition af datasæt

**Comentado [KM21]:** Student helpers: If possible, please translate this blue box and the following boxes.


**Screening og udvælgelse af kandidater: definition af datasæt**

Mens virksomheden undersøger alle aspekter og muligheder inden for AI-rekruttering, så vurderer den også AI nøje i forhold til screenings- og udvælgelsesfasen. Dette indebærer en dybdegående analyse af CV-scanning og at lave et passende match mellem kandidater og de ledige stillinger.

Ved at anvende AI-værktøjer til at automatisere CV-screeningsprocessen er målet hurtigt og præcist at finde frem til de kandidater, der opfylder jobkravene. Anvendelsen af denne teknologi forventes at føre til en mere strategisk brug af HR-ressourcer, forbedre kvaliteten af de kandidater, der shortlistes, og øge den samlede effektivitet i ansættelsesprocessen ved at afstemme kandidaternes færdigheder og baggrund med virksomhedens krav.

Først skal virksomheden definere de datasæt, som er grundlaget for rekrutteringssystemet, og som algoritmerne henter data fra.

## Træning af AI

Træningsdatasæt spiller en afgørende rolle i udviklingen af algoritmer. Disse datasæt er det fundament, som algoritmerne er bygget på. Formålet med disse datasæt er at fungere som "grundsandheden" eller det faktuelle grundlag, der instruerer algoritmen i, hvordan den skal fortolke og behandle store mængder data. De lærer algoritmen at forstå forholdet mellem inputvariabler (de oplysninger, der tilføres systemet) og en outputvariabel (den forudsigelse eller beslutning, som systemet træffer på baggrund af input). Der opstår en betydelig udfordring, når selve træningsdataene indeholder bias. Disse bias kan være en afspejling af historiske uligheder, fordomme eller simpelthen resultatet af en ikke-repræsentativ eller ufuldstændig dataindsamling. Når en algoritme trænes på biased data, lærer den disse bias og viderefører dem i sine forudsigelser og beslutninger. Dette fænomen beskrives almindeligvis som "*bias in, bias out*"-problemet. Det betyder, at hvis inputdataene (*bias in*) er biased, vil maskinlæringsmodellens output (*bias out*) også være biased. Det kan føre til uretfærdige, diskriminerende eller fejlbehæftede resultater, hvor kandidater udvælges eller afvises på baggrund af biased kriterier i stedet for deres sande kvalifikationer eller potentiale (Sheard, 2022).

## Træning af AI på virksomhedens interne data



Når virksomheder træner AI-applikationer til CV-screening, har de adgang til en række interne datakilder, der kan hjælpe i læringsprocessen (Fernandes, 2021; Ma, 2024; Sheard, 2022).

En sådan kilde er data om medarbejdernes performance kombineret med historiske ansættelsesdata, som omfatter interne optegnelser over medarbejdervurderinger og tidligere ansættelsesresultater. Derudover kan virksomheder bruge en akkumulering af indsendte CV'er gennem tiden, som omfatter et væld af historiske CV'er, der er modtaget i forbindelse med jobansøgninger.

En anden værdifuld ressource er data om "falske negative eksempler", som er oplysninger om kandidater, der fejlagtigt er blevet afvist i tidligere rekrutteringer. Analyse af disse sager kan give indsigt i mønstre, som AI'en bør undgå at gentage. Desuden kan virksomheder udvælge specifikke nøgleord og sætninger, der ofte forekommer i de indsamlede CV'er, og bruge dem til at forfine AI'ens evne til at identificere relevante kvalifikationer og erfaringer.

Endelig kan oplysninger om uddannelsesmæssig baggrund, som er en standardkomponent i ansøgninger, også udnyttes til at forbedre AI'ens forståelse af akademiske kvalifikationer og deres relevans for forskellige jobfunktioner.

### Træning af AI på virksomhedens eksterne data

Mulige kilder til eksterne data i AI-træning er blevet foreslået gennem tiden og omfatter flere platforme og metoder til at indsamle information (Banerjee, 2022). Sociale medieplatforme er en rig kilde, hvor enkeltpersoner deler personlige interesser, erfaringer og opnåede, faglige resultater, hvilket giver indsigt i en kandidats personlighed, færdigheder og professionelle netværk (Albaroudi et al., 2024; Albassam, 2023; Chaker, 2018; Filtered, n.d.; Heymans, 2022; Shipman, 2021). Derudover indeholder professionelle hjemmesider, der er designet til netværk og karriereudvikling, såsom LinkedIn, detaljerede profiler, der inkluderer arbejdshistorik, uddannelse, færdigheder, anbefalinger og opnåede, faglige resultater. Disse profiler giver et omfattende overblik over potentielle talenter. Endelig kan internetsøgninger foretaget af virksomheder finde offentligt tilgængelige oplysninger om en kandidat fra kilder som nyhedsartikler, personlige blogs, publikationer eller andre websider. Denne brede vifte af oplysninger giver en værdifuld kontekst om en persons kvalifikationer, præstationer og karakter (Chaker, 2018).

## 7. Udvikling af den algoritmiske model




**Jobannonce – udvælgelse af nøgleord til algoritmen**

Virksomheden har for nylig offentliggjort en jobbeskrivelse og modtaget hundredvis af ansøgninger.

Prototypen til AI-screening og shortlisting scanner ansøgernes CV'er for nøgleord og andre oplysninger, der menes at være relevante for en vellykket ansættelse, f.eks. erfaring, jobtitler, tidligere arbejdsgivere, universiteter og universitetsgrader. Baseret på dette opretter systemet en struktureret ansøgerprofil, og alle kandidater får derefter en score og bliver rangeret (Sheard, 2022).

AI strømliner i høj grad screeningen af ansøgere, som traditionelt er et af de mest tidskrævende trin i rekrutteringen. I screeningsfasen af 'rekrutteringstragten' er vurderingen af jobkandidaterne baseret på deres erfaring, færdigheder og andre relevante egenskaber afgørende for at udvælge dem, der inviteres til samtale. AI kan bruges til at scanne CV'er for nøgleord og andre data, der er forbundet med

vellykkede ansættelser – såsom jobtitler, tidligere arbejdsgivere, uddannelsesinstitutioner og kvalificationer – og dermed strømline processen med at identificere egnede kandidater (Dennis, 2023; Fernandes, 2021; Sheard, 2022).

Ved at bruge AI-værktøjer kan rekrutteringsansvarlige hurtigt identificere og prioritere de mest relevante kandidater og dermed øge effektiviteten og forbedre vigtige rekrutteringsmetrikker som f.eks. *'time-to-hire'*, dvs. hele den tid det tager at ansætte en kandidat. AI-systemer frasorterer automatisk kandidater, der ikke opfylder specifikke foruddefinerede kriterier, analyserer CV'er for at fremhæve relevant erfaring og rangordner ansøgere i henhold til deres egnethed til jobbet. Denne automatisering gør det muligt for rekrutteringsmedarbejdere at fokusere mere på at evaluere kandidater i de senere faser af rekrutteringstragten. Derudover kan AI scanne CV'er i den indledende fase af CV-screening og udpege kandidater, der opfylder specifikke kriterier. Maskinlæringsalgoritmer vurderer derefter disse kandidaters færdigheder og kvalifikationer i forhold til jobspecifikationerne for at udarbejde en liste over egnede ansøgere (ResumeMent, 2023; Dennis, 2023).

Jobbeskrivelsen indeholder vigtige detaljer om rollen, herunder nødvendige færdigheder, kvalifikationer og erfaring, som AI'en bruger til at screene CV'erne. Uden en detaljeret jobbeskrivelse ville AI'en mangle de nødvendige parametre for, hvad den skal kigge efter i CV'erne.

152

**Comentado [KM22]:** Nidhi: Should this be translated? If yes, how can I/we edit the figure?

**Job description for Communication Assistant**
We are looking for a communications assistant to be responsible for the creation of content such as media releases, blogs, and social media posts on behalf of our company. You will also be monitoring media and campaign coverage and attending internal and external events.

**Responsibilities**
- Develop and manage diverse content, including media releases, blogs, and social media posts. aligned with the company's strategic goals; monitor media and campaign coverage.
- Support the implementation of internal and external communications strategies and assist in managing the company's image.
- Organize marketing events and provide comprehensive administrative support. including maintaining event calendars and updating contact lists.
- Compile and prepare presentations and reports while tracking project progress and media exposure.

**Requirements**
- Holds a Bachelor's degree in communications. marketing, or related field, with excellent verbal and written communication abilities.
- Proficient in social media strategies and media relations, with a creative and innovative approach.
- Strong organizational skills and attention to detail, capable of multitasking and maintaining positive interpersonal relationships.
- Skilled in using office management and design software, such as Photoshop and InDesign, and knowledgeable in various social media platforms.

Fremskridt inden for AI har bevæget sig ud langt over simpel matchning af søgeord. Oprindeligt brugte virksomheder algoritmer til at scanne CV'er for udvalgte ord eller sætninger. Nutidens AI-teknologier, herunder chatbots og

sofistikerede CV-fortolkningsværktøjer, vurderer nu semantisk relevans og relaterede termer for at få et mere præcist billede af en kandidats kvalifikationer. Derudover anvendes maskinlæringsalgoritmer til at forudsige fremtidige jobpræstationer baseret på forskellige indikatorer, f.eks. arbejdstid og produktivitet. Disse værktøjer kan også foreslå de bedst egnede ledige stillinger til kandidater baseret på deres profiler (Hunkenschroer & Luetge, 2022).

Men i dette kursus bruger vi simpel søgeordsmatchning som eksempel for at illustrere, hvordan en AI CV-screening-applikation behandler ansøgninger.

**Liste over specifikke søgeord, der bruges til screening af kandidater**

- "Bachelorgrad i kommunikation"
- "Bachelorgrad i marketing" "Skabelse af indhold"
- "Medieudgivelser"
- "Blogs"
- "Indlæg på sociale medier"
- "Overvågning af medier" og "kampagnedækning"
- "Støtte til implementering af kommunikationsstrategier"
- "Håndtering af virksomhedens image"
- "Planlægning af markedsføringsevents"
- "Yde omfattende administrativ støtte" "Opdatere lister på kontakter"
- "Udarbejdelse af præsentationer" og "rapporter" "Holde styr på projektets fremskridt" og "medieeksponering"
- "Dygtig til strategier for sociale medier" og "medierelationer"
- "Kreativ og innovativ tilgang"
- "Stærke organisatoriske evner" og "detaljeorienteret"
- "I stand til at multitaske"
- "Opretholdelse af positive personlige relationer" "Dygtig til kontorstyring" og "designsoftware"
- "Photoshop" og "InDesign"
- "Viden om forskellige sociale medieplatforme"



Ifølge Sheard (2022) gennemgås CV'er typisk kun af menneskelige øjne, hvis de stemmer overens med jobannoncens kriterier, mens resten kan ignoreres, hvilket understreger et potentielt problem, hvor en betydelig del af CV'erne kan blive afvist for tidligt.

Nu vil vi analysere to ansøgninger for at forstå, hvordan prototypen på AI-applikationen vurderer kandidater i screeningsfasen. Denne analyse vil hjælpe med at identificere, om der er nogen bias, der påvirker udvælgelsesprocessen.


**Sammenligning af to ansøgeres CV'er med henblik på formuleringen af søgeord**

Virksomheden arrangerer en workshop for at se nærmere på AI-drevet CV-screening. Til formålet bruger de et casestudie, baseret på virksomhedens egne aktiviteter.

I løbet af workshoppen analyserer de to CV'er parallet for at afdække, hvordan forskelle i ordvalg og formuleringer utilsigtet kan føre til, at man overser kvalificerede kandidater.


JOHN DOE

Communications assistant

About Me

A dedicated communications professional with a Bachelor's degree in Marketing, seeking the role of Communications Assistant to leverage extensive experience in content creation, social media management, and event coordination to contribute to the company's strategic goals.

+123-456-7890

hello@doeland.com

123 Anywhere St., Any City

EXPERIENCE

**Communications assitant**
**ABC Corporation**
**2019- Present**

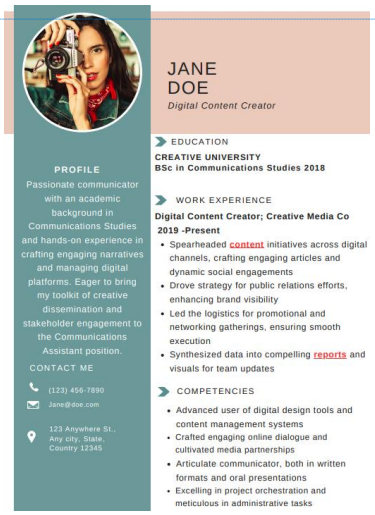
- Developed and managed diverse content, including media releases, blogs, and social media posts
- Supported the implementation of comprehensive communications strategies
- Organized multiple marketing events and maintained event calendars
- Compiled presentations and tracked project progress, ensuring alignment with strategic goals

EDUCATION

**University of communications**
Bachelor's degree in Marketing
2014-2018

SKILLS SUMMARY

- Proficient in Photoshop and InDesign
- Skilled in social media strategies and media relations
- Excellent verbal and written communication abilities
- Strong organizational skills and attention to detail


JANE DOE

*Digital Content Creator*

PROFILE

Passionate communicator with an academic background in Communications Studies and hands-on experience in crafting engaging narratives and managing digital platforms. Eager to bring my toolkit of creative dissemination and stakeholder engagement to the Communications Assistant position.

CONTACT ME

(123) 456-7890

Jane@doe.com

123 Anywhere St., Any city, State, Country 12345

EDUCATION

**CREATIVE UNIVERSITY**
**BSc in Communications Studies 2018**

WORK EXPERIENCE

**Digital Content Creator; Creative Media Co**
**2019 -Present**

- Spearheaded content initiatives across digital channels, crafting engaging articles and dynamic social engagements
- Drove strategy for public relations efforts, enhancing brand visibility
- Led the logistics for promotional and networking gatherings, ensuring smooth execution
- Synthesized data into compelling reports and visuals for team updates

COMPETENCIES

- Advanced user of digital design tools and content management systems
- Crafted engaging online dialogue and cultivated media partnerships
- Articulate communicator, both in written formats and oral presentations
- Excelling in project orchestration and meticulous in administrative tasks

**Comentado [KM23]:** Nidhi: Should the resumes be translated??


PROFILE

Passionate communicator with an academic background in Communications Studies and hands-on experience in crafting engaging narratives

my toolkit of creative dissemination and stakeholder engagement to the Communications Assistant position.

CONTACT ME

(123) 456-7890

Jane@doe.com

123 Anywhere St., Any city, State, Country 12345

154

Dette eksempel illustrerer effekten af forskelle i ordlyd og formuleringer i to CV'er, og hvordan de kan påvirke resultatet af en AI-screeningsproces. Selv om det andet CV har relevante kvalifikationer, kan det blive frasorteret, fordi beskrivelserne ikke stemmer overens med de specifikke nøgleord, som virksomheden har fastlagt for screeningen. Dette viser, hvor vigtigt det er at matche sproget i et CV med de nøgleord, som et AI-system forventer.

| Aspect | CV1 | CV 2 |
|---|---|---|
| Job Titles and Terminology | Uses standard terms like "Communications Assistant" | Uses titles like "Digital Content Creator" |
| Academic Background | Explicitly mentions "Bachelor's Degree in Marketing" | States "BSc in Communications Studies" |
| Describing Tasks and Skills | Direct match with keywords like "Content creation", "Media releases" | Uses varied language like "crafting engaging narratives", "spearheaded content initiatives" |
| Technical and Software Skills | Specifies "Photoshop" and "InDesign" | General mention of "digital design tools and content management systems" |
| Direct Keyword Matches | Closely matches many specified keywords | Uses different phrasing that may not match the specific Keywords set for screening |

**Comentado [KM24]:** Nidhi: Should this table be translated? If yes, do you have any idea how?

155

## 8. Hvordan man måler retfærdighed i maskinlæringsløsninger

**Comentado [KM25]:** A math-person should quality check pp 155-160 to see that math and statistical expressions are correct.


**Test af løsningens retfærdighed**

En evaluering af algoritmeprototypen afslørede uoverensstemmelser i CV-analysen, hvilket gav anledning til forbedringer for at opnå en mere omfattende forståelse af CV-attributter. Modellen er nu klar til en testfase, inden den tages i brug i den virkelige verden.

Selvom en evaluering af flere aspekter er påkrævet, vil fokus her være på at teste applikationens retfærdighed for at vurdere omfanget af eventuelle bias, der opstår i processen. Fairness-metrikker anvendes til at måle niveauet af bias i applikationens beslutninger.

For at skabe systemer, der er retfærdige for alle og fri for bias såsom køns- eller racediskrimination, er det vigtigt at evaluere løsningens performance på tværs af

forskellige befolkningssegmenter. Der findes forskellige metoder til at vurdere en løsnings retfærdighed, men i denne sammenhæng vil vi udelukkende fokusere på binære klassifikationstests *(Machine Learning and Fairness*, 2021).

Retfærdighed er blevet et vigtigt emne inden for maskinlæring på grund af den stigende anvendelse på forskellige områder. Omfattende forskning har fokuseret på dette område, fordi maskinlæringssystemer er stærkt afhængige af data, som, når de stammer fra mennesker, i sagens natur kan være forudindtagede. Det har vist sig at være en udfordring at definere retfærdighed i matematiske termer, hvilket har ført til manglende konsensus om standardformuleringer. Men som Zhong (2018) bemærker, samler de fleste definitioner af retfærdighed sig om flere nøglebegreber (nedenstående oversættelser i tillempet danske versioner):

- Uvidenhed (*unawareness*)
- Demografisk paritet (*demographic parity*)
- Udlignede odds *(equalized odds)*
- Forudsigende paritet (*predicative rate parity*)
- Individuel retfærdighed (*individual fairness*)
- Kontrafaktisk retfærdighed (*counterfactual fairness*)

Demografisk paritet, udlignede odds og forudsigende paritet grupperes typisk under paraplyen "grupperetfærdighed" (Zhong, 2018). Dette vil blive beskrevet nærmere i dette afsnit, især set i lyset af AI i rekruttering.

Ud over grupperetfærdighed opdeler Castelnovo et al. (2022) retfærdighedskategorierne i individuel retfærdighed og kausalitetsbaseret retfærdighed, hvilket også omfatter de emner, der er nævnt i definitionerne i Zhong (2018).



## Forvirringsmatrix

En af de binære klassificeringsmetoder er forvirringsmatricen *(Machine Learning and Fairness*, 2021). For at forstå retfærdighedsmålinger starter vi med begrebet forvirringsmatrix. Dette værktøj opsummerer forudsigelserne fra en model i forhold til de faktiske resultater, eller grundsandheden, som den blev trænet på. Forvirringsmatrixen viser antallet af både korrekte og forkerte forudsigelser, hvilket giver et klart billede af modellens performance og dens forklarlige validitet (*explainable validity*) (Saplicki, 2022).

| | Unqualified | Qualified |
|---|---|---|
| Reject | True Negative (TN) | False Negative (FN) |
| Hire | False Positive (FP) | True Positive (TP) |

BILLEDKILDE | Billede 1 - Forvirringsmatrix (Modificeret fra Machine Learning and Fairness, 2021)

**Comentado [KM26]:** To translate or not to translate the following tables?

En forvirringsmatrix hjælper os med at kategorisere resultaterne af ansættelsesprocessen. Kvalificerede kandidater, der ansættes med succes, betragtes som sande positive (TP), og ukvalificerede kandidater, der afvises korrekt, er sande negative (TN). Omvendt henviser falske positive (FP) til ukvalificerede kandidater, der fejlagtigt ansættes, mens falske negative (FN) er kvalificerede kandidater, der fejlagtigt afvises. En forvirringsmatrix kategoriserer simpelthen ansøgere i forskellige grupper baseret på deres klassificeringsresultater. De fire værdier, der fremkommer i en forvirringsmatrix, bruges til at beregne standard maskinlæringsmålinger, herunder nøjagtighed og præcision, på tværs af forskellige demografiske grupper (*Machine Learning and Fairness*, 2021; Teodorescu, 2020).

I det rekrutteringsscenarie, vi ser på her, vil vi fokusere specifikt på kønsdemografi. Selvom det er vigtigt at overveje andre demografiske faktorer, vil dette eksempel udelukkende koncentrere sig om køn.

Lad os antage, at der er 100 hhv. mandlige og kvindelige ansøgere. Systemet har nøjagtigt identificeret 75 personer fra hver gruppe som enten kvalificerede eller ukvalificerede, hvilket viser, at graden af nøjagtighed er konsistent på tværs af begge demografier (*Machine Learning and Fairness*, 2021).

| MEN | Unqualified | Qualified | WOMEN | Unqualified | Qualified |
|---|---|---|---|---|---|
| Reject | **15 (TN)** | 5 (FN) | Reject | **60 (TN)** | 20 (FN) |
| Hire | 20 (FP) | **60 (TP)** | Hire | 5 (FP) | **15 (TP)** |

BILLEDKILDE | Billede 2 - Eksempel på forvirringsmatrix i forbindelse med rekruttering (modificeret fra Machine Learning and Fairness, 2021)

## Demografisk paritet

Med udgangspunkt i de leverede data vil vi undersøge, hvordan forskellige retfærdighedsmålinger anvendes i denne sammenhæng, begyndende med demografisk paritet. Denne metrik tager kun højde for klassificeringsoutput uden hensyn til de faktiske betegnelser. Demografisk paritet er en af de foreslåede metrikker, når man vil evaluere et automatiseret ansættelsessystem (*Machine Learning and Fairness*, 2021).

Demografisk paritet, almindeligvis kendt som uafhængighed eller statistisk paritet, er et kriterium for retfærdighed, hvor sandsynligheden for et gunstigt resultat, såsom at blive ansat, ikke er påvirket af en beskyttet egenskab som køn (Zhong, 2018).

| MEN | Unqualified | Qualified |
|---|---|---|
| Reject | 15 (TN) | 5 (FN) |
| Hire | **20 (FP)** | **60 (TP)** |

| WOMEN | Unqualified | Qualified |
|---|---|---|
| Reject | 60 (TN) | 20 (FN) |
| Hire | **5 (FP)** | **15 (TP)** |

BILLEDKILDE | Billede 3 - Eksempel på demografisk paritet i rekrutteringsproces (modificeret fra Machine Learning and Fairness, 2021)

Hvis 80 ud af 100 mandlige ansøgere og kun 20 ud af 100 kvindelige ansøgere bliver ansat i rekrutteringsscenariet (billede 3), er den demografiske paritet ikke opfyldt. Det kan dog være acceptabelt, hvis de mandlige ansøgere faktisk er mere kvalificerede end de kvindelige ansøgere, som det ser ud til i dette tilfælde (*Machine Learning and Fairness*, 2021).

Demografisk paritet er dog svært at opnå, idet enkeltpersoner kan tilhøre flere beskyttede grupper, da lige sandsynligheder måske ikke er mulige på tværs af alle demografier. Demografisk paritet kan nogle gange utilsigtet favorisere mindre kvalificerede personer på grund af dens fokus på gruppeniveau, hvilket potentielt kan føre til uretfærdighed på individniveau. Hvis man f.eks. øger andelen af mindre kvalificerede kvindelige ansøgere, kan det mindske chancerne for at ansætte kvalificerede mandlige ansøgere (Teodorescu, 2020). Derfor tager dette mål kun hensyn til resultatet, ikke de sande betegnelser, hvilket betyder, at det ikke tager højde for ansøgernes faktiske kvalifikationer. Den siger, at forholdet mellem ansatte ansøger og det samlede antal ansøgere skal være ens for alle (*Machine Learning and Fairness*, 2021).



## Forudsigende paritet

Tilstrækkelighed, set fra perspektivet af dem, der modtager den samme beslutning fra en model, kræver lige store fejlrater uanset den eventuelle forekomst af følsomme attributter. Dette koncept er omfattet af forudsigende paritet (*predictive parity*), også kendt som udfaldstesten. Den sikrer, at fejlprocenterne er afbalancerede for personer, der er grupperet efter de afgørelser, de modtager, snarere end efter de faktiske resultater. I modsætning til demografisk paritet tager forudsigende paritet både klassifikatorens afgørelser og de sande resultater i betragtning. Den vurderer, om sandsynligheden for, at en ansøger er egnet til en stilling, er konsistent på tværs af forskellige grupper, forudsat at AI'en har udvalgt dem til ansættelse. Forventningen er, at denne sandsynlighed skal være næsten identisk for alle de grupper, der betragtes (Castelnovo et al., 2022; *Machine Learning and Fairness*, 2021).

| MEN | Unqualified | Qualified | WOMEN | Unqualified | Qualified |
|---|---|---|---|---|---|
| Reject | 15 | 5 | Reject | 60 | 20 |
| Hire | 20 | **60** | Hire | 5 | **15** |

BILLEDKILDE | Billede 4 - Eksempel på forudsigende paritet i rekrutteringsproces (Modificeret *fra Machine Learning and Fairness*, 2021)

I vores rekrutteringseksempel opfylder systemet kriterierne for forudsigende paritet, da tre fjerdedele af de ansatte ansøgere fra begge grupper faktisk er kvalificerede *(Machine Learning and Fairness*, 2021).

## Falsk positiv rate-balance, falsk negativ rate-balance og udlignede odds

Vedr. målingen, der er kendt som den **falsk positiv rate**, hævdes det, at sandsynligheden for, at en ukvalificeret kandidat fejlagtigt bliver ansat, bør være konsekvent på tværs af alle demografiske forhold. Sagt på en anden måde skal systemet i lige høj grad fejlvurdere ukvalificerede kandidater, uanset deres køn (*Machine Learning and Fairness*, 2021).

**False positive rate**

| MEN | Unqualified | Qualified |
|---|---|---|
| Reject | 15 | 5 |
| Hire | **20** | 60 |

| WOMEN | Unqualified | Qualified |
|---|---|---|
| Reject | 60 | 20 |
| Hire | **5** | 15 |



BILLEDKILDE | Billede 5 - Eksempel på falsk positiv rate i et rekrutteringsforløb (Modificeret *fra Machine Learning and Fairness*, 2021)

I dette eksempel (billede 5) er andelen af falsk positive anderledes, idet en betydeligt større andel af ukvalificerede mænd bliver ansat end ukvalificerede kvinder.

**Falsk negativ** rate-balance er stort set det samme som falsk positiv rate, men blot for falsk negativ rate. Begrebet paritet for falsk negativ rate antyder, at sandsynligheden for fejlagtigt at afvise kvalificerede kandidater bør være ensartet på tværs af alle grupper.

**False negative rate**

| MEN | Unqualified | Qualified |
|---|---|---|
| Reject | 15 | **5** |
| Hire | 20 | 60 |

| WOMEN | Unqualified | Qualified |
|---|---|---|
| Reject | 60 | **20** |
| Hire | 5 | 15 |

BILLEDKILDE | Billede 6. Eksempel på falsk negativ rate i et rekrutteringsforløb (Modificeret *fra Machine Learning and Fairness*, 2021)

I dette tilfælde (billede 6) er der en bemærkelsesværdig forskel i falsk negative rater, hvor kvalificerede kvindelige kandidater bliver afvist i højere grad end deres mandlige modparter.

**Udlignede odds** (*equalized odds*) kombinerer disse to egenskaber ved at opfylde både falsk positiv rate-balance og falsk negativ rate-balance, hvilket betyder, at systemet skal håndtere begge typer fejl – dvs. fejlagtigt accepterede ukvalificerede kandidater og fejlagtigt afviste kvalificerede kandidater – på samme måde og på tværs af forskellige grupper, hvilket sikrer, at alle kandidater med samme kvalifikationer får en ensartet behandling (*Machine Learning and Fairness*, 2021). Udlignede odds er svære at opnå, men når de opnås, kan de betragtes som et af de højeste niveauer af algoritmisk retfærdighed (Teodorescu, 2020).

## Afvejning af metrikker

Et system kan ikke på samme tid opfylde forudsigende paritet, balance i falsk positiv rate og balance i falsk negativ rate på grund af matematiske begrænsninger. Hvis et system opfylder to af disse kriterier, vil det uundgåeligt komme til kort i forhold til det tredje (*Machine Learning and Fairness*, 2021).

Dette understreger de uundgåelige afvejninger mellem forskellige retfærdighedsmål. Når vi stræber efter retfærdighed, støder vi ofte på modstridende mål, og det er stadig svært at opnå perfekt balance på tværs af alle dimensioner.

I praksis skal beslutningstagerne nøje afveje disse kompromiser, se på konteksten og træffe informerede valg baseret på applikationens specifikke behov og formål. Selvom perfektion måske er uopnåelig, er en kontinuerlig indsats for retfærdighed og gennemsigtighed afgørende i algoritmiske beslutningssystemer.

Endelig er det vigtigt at erkende, at mange aspekter af retfærdighed ikke kan kvantificeres på grund af deres sociotekniske kompleksitet. Derfor er det ikke alle aspekter af retfærdighed, der kan sammenfattes i metrikker (*Machine Learning and Fairness* , 2021).

## Individuel retfærdighed

Individuel retfærdighed adskiller sig fra gruppebaserede retfærdighedskriterier (de tidligere omtalte metrikker) ved at fokusere på behandlingen af enkeltpersoner. I dette koncept hævdes det, at individer med lignende karakteristika bør få sammenlignelige resultater. Det er komplekst at finde en metode til at måle individuelle ligheder. Forestil dig tre jobkandidater: A med en bachelorgrad og et års relevant erfaring; B med en kandidatgrad og samme mængde erfaring; og C med en kandidatgrad, men ingen erfaring. At afgøre, om

A ligner B eller C mest, og at kvantificere denne lighed, er en udfordring uden at kunne sammenligne deres jobpræstation, hvilket ikke er muligt, hvis ikke alle tre ansættes (Zhong, 2018).

## 9. Implementering af modellen


**Informér ansøgere om brug af AI i rekrutteringen**

Efter omhyggelige interne evalueringer er virksomheden klar til at lancere sit screeningsværktøj på nogle bestemte stillingsopslag for at indsamle den første feedback fra brugere og ansøgere.

Det er afgørende, at ansøgerne hele tiden er klar over, at de er i kontakt med et AI-drevet rekrutteringssystem, så tilliden dermed fremmes. Der bør skabes opmærksomhed og bevidsthed om fordelene ved AI-værktøjer helt fra starten, hvilket øger ansøgernes villighed til at bruge interaktiv teknologi som chatbots. Derudover er det vigtigt kontinuerligt at forbedre AI-rekrutteringsprocesserne og sikre klarhed om systemets funktion, da dette er afgørende for, at det bliver taget i brug og bliver en god hjælp for ansøgerne.

Det skal også sikres, at der er et menneske, der fører tilsyn og griber ind, når det er nødvendigt (Chen, 2023).

## 10. Drift og overvågning


**Efter udrulning**

Mens interne vurderinger af retfærdighed blev gennemført før lanceringen, skal virksomheden også være klar til uafhængige eksterne revisioner.

Det er også en fordel løbende at efterspørge feedback efter implementeringen og etablere kanaler, hvor igennem man kan holde kontakten til interessenter. Dette indebærer at holde medarbejderne informeret og involveret i høringer og deltagelsesprocesser under hele implementeringen af AI-systemet i organisationen (HLEG Trustworthy AI).

### Revision af AI



Mens interne retfærdighedsmålinger foretages før implementeringen, er en ekstern revision også noget, som virksomheden skal være forberedt på.

Ifølge Ahmed (2020) bør revisorer, der evaluerer AI-applikationer, prioritere to aspekter: at sikre overholdelse af de registreredes rettigheder og vurdere teknologiske risici inden for maskinlæring og cybersikkerhed. Revisionen begynder med at definere dens omfang, mål og de AI-relaterede risici for organisationen, typisk dokumenteret i en risiko- og kontrolmatrix inden for nogle etablerede rammer.

De vigtigste risici er forkert tilpasning af IT-strategier til virksomhedsmålene, ineffektiv styring og uklare ansvarsstrukturer. Med hensyn til *compliance* skal revisorer forstå principperne for databeskyttelse og deres konsekvenser for enkeltpersoners rettigheder i henhold til regler som GDPR (Ahmed, 2020).

Der findes flere retningslinjer for revision af AI. Efterhånden som AI omformer erhvervslivet og samfundet, skal revisorer sikre, at de er klar til at tackle nye udfordringer og verificere effektiviteten af kontroller og styring omkring AI (Ahmed, 2020).

## 11. Konklusion

I denne kursusdel var der fokus på at undersøge algoritmisk bias i en praktisk kontekst. Der blev set nærmere på emnet gennem et hypotetisk scenarie, hvor en international virksomhed udvikler en AI-løsning til rekruttering. Når AI-teknologier integreres i rekruttering, giver de betydelige fordele ved at forbedre effektiviteten og potentielt øge mangfoldigheden på arbejdspladsen ved at lægge vægt på kvalifikationer frem for demografiske karakteristika. Men implementeringen af disse teknologier rejser også betydelige etiske udfordringer, især når det gælder om at identificere og afbøde effekterne af algoritmisk bias.

AI-systemer kan utilsigtet videreføre eksisterende bias, der findes i deres træningsdata. Disse bias kan afspejle historiske forskelle eller samfundsmæssige uligheder, hvilket fører til diskriminerende praksis i rekrutteringsprocesser. Det er ikke kun et etisk krav, men også en juridisk nødvendighed at tage hånd om disse bias, da lovgivningen om AI i rekruttering er i konstant udvikling. Organisationer skal være årvågne og proaktive i forhold til at overholde disse standarder for at forhindre diskrimination.

For effektivt at afbøde konsekvenserne af disse udfordringer er det afgørende at bruge sofistikerede værktøjer, der er specielt designet til at identificere og korrigere bias i AI-systemer. Regelmæssig overvågning og uafhængige revisioner er afgørende, da de hjælper organisationer med at identificere bias og justere deres AI-systemer i overensstemmelse hermed.



Desuden er det afgørende at opretholde en balance mellem AI-drevne processer og menneskeligt tilsyn. Mens AI kan strømline rekrutteringsprocesser betydeligt, er det menneskelige element stadig uundværligt til at fortolke sammenhænge, som AI måske fejlvurderer. Menneskeligt tilsyn sikrer, at rekruttering forbliver en hensynsfuld og inkluderende proces, der respekterer nuancerne i menneskelige erfaringer og værdier.

Sammenfattende kan man sige, at selvom AI i rekruttering giver betydelige fordele, er det afgørende at udnytte denne teknologi på en ansvarlig måde. Organisationer bør fokusere på løbende at identificere bias og anvende avancerede værktøjer til at afbøde dem. Denne afbalancerede tilgang sikrer, at rekrutteringsprocesserne opretholder de højeste standarder for retfærdighed og etisk praksis, hvilket gør det muligt for virksomheder at tage teknologiske fremskridt til sig og samtidig sikre mangfoldighed og retfærdighed på arbejdspladsen.

## 12. Referencer


Ahmed, H. S. A. (2020, December 21). *Auditing Guidelines for Artificial Intelligence*. ISACA. https://www.isaca.org/resources/news-and-trends/newsletters/atisaca/2020/volume-26/auditing-guidelines-for-artificial-intelligence

Alan Turing Institute. (2022). *Reflect on Purpose, Positionality, and Power—Turing Commons*. https://alan-turing-institute.github.io/turing-commons/skills-tracks/aeg/chapter5/reflect/

Albaroudi, E., Mansouri, T., & Alameer, A. (2024). A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring. *AI*, *5*(1), 383–404. https://doi.org/10.3390/ai5010019

Albassam, W. A. (2023). The Power of Artificial Intelligence in Recruitment: An Analytical Review of Current AI-Based Recruitment Strategies. *International Journal of Professional Business Review*, *8*(6), e02089. https://doi.org/10.26668/businessreview/2023.v8i6.2089

ALTAI. (2020). *The Assessment List for Trustworthy Artificial Intelligence*. https://altai.insight-centre.org/

Arivu Recruitment and Consulting. (2023, July 13). *Pros and Cons of Using Artificial Intelligence (AI) in Recruiting | LinkedIn*. https://www.linkedin.com/pulse/pros-cons-using-artificial-intelligence/

Banerjee, S. (2022, February 20). *Big data in Recruitment Sector | LinkedIn*. https://www.linkedin.com/pulse/big-data-recruitment-sector-shantanu-banerjee-phd/

Bursell, M., & Roumbanis, L. (2024). After the algorithms: A study of meta-algorithmic judgments and diversity in the hiring process at a large multisite company. *Big Data & Society*, *11*(1), 20539517231221758. https://doi.org/10.1177/20539517231221758

CareerExperts. (2023, April 24). Reducing Bias in Hiring with AI-Powered Job Description Generation. *Career Experts*. https://www.careerexperts.co.uk/management-leadership/ai-powered-job-description

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, *12*(1), 4209. https://doi.org/10.1038/s41598-022-07939-1

Chaker, N. (2018, October 24). *Candidate Data Sources: The Recruiter's Checklist*. https://beamery.com/resources/blogs/candidate-data-sources-the-recruiter-s-checklist

Chen, Z. (2023). Collaboration among recruiters and artificial intelligence: Removing human prejudices in employment. *Cognition, Technology & Work*, *25*(1), 135–149. https://doi.org/10.1007/s10111-022-00716-0

Dennis, J. (2023, May 22). *AI Recruiting: Uses, Advantages, & Disadvantages 2024*. TechnologyAdvice. https://technologyadvice.com/blog/human-resources/ai-recruiting/

DigitalOcean. (n.d.). *How to Use AI in Hiring: Techniques and Tools*. Retrieved April 26, 2024, from https://www.digitalocean.com/resources/article/ai-in-hiring

Featured. (2023, December 22). *How these 6 leaders are using AI-powered onboarding*. Fast Company. https://www.fastcompany.com/90996367/how-leaders-using-ai-powered-onboarding

Fernandes, R. F. (2021). Big Data as a Tool to Enhance Recruitment Processes. *E-Journal of International and Comparative LABOUR STUDIES*, *10*(03).

Filtered. (n.d.). *What is AI Resume Screening and How Can it Benefit Your Enterprise? | Filtered Blog*. Retrieved March 14, 2024, from https://www.filtered.ai/blog/ai-resume-screening-fd

Friedman, B., & Nissenbaum, H. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems*, *14*(3), 330–347.

Fritts, M., & Cabrera, F. (2021). AI recruitment algorithms and the dehumanization problem. *Ethics and Information Technology*, *23*(4), 791–801. https://doi.org/10.1007/s10676-021-09615-w

Giordano, F., Morelli, N., Götzen, A. D., & Hunziker, J. (2018, June). *The stakeholder map: A conversation tool for designing people-led public services*. ServDes2018 - Service Design Proof of Concept, Politecnico di Milano.

Google for Developers. (2022). *Fairness: Types of Bias | Machine Learning*. Google for Developers. https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias

Harvard John A. Paulson School of Engineering and Applied Sciences. (2023, December 6). *How Can Bias Be Removed from Artificial Intelligence-Powered Hiring Platforms?* https://seas.harvard.edu/news/2023/06/how-can-bias-be-removed-artificial-intelligence-powered-hiring-platforms

Harwell, D. (2019a, July 7). FBI, ICE find state driver's license photos are a gold mine for facial-recognition searches. *Washington Post*. https://www.washingtonpost.com/technology/2019/07/07/fbi-ice-find-state-drivers-license-photos-are-gold-mine-facial-recognition-searches/

Harwell, D. (2019b, December 19). Federal study confirms racial bias of many facial-recognition systems, casts doubt on their expanding use. *Washington Post*. https://www.washingtonpost.com/technology/2019/12/19/federal-study-confirms-racial-bias-many-facial-recognition-systems-casts-doubt-their-expanding-use/

Heymans, Y. (2022, September 7). *Machine learning in recruitment: A deep dive*. https://www.herohunt.ai/blog/machine-learning-in-recruitment-a-deep-dive

Hidalgo, M. A. S. (2019). *Design of an Ethical Toolkit for the Development of AI Applications* [Delft University of Technology]. http://resolver.tudelft.nl/uuid:b5679758-343d-4437-b202-86b3c5cef6aa

Hoory, L., & Botorff, C. (2022, August 7). *What Is A Stakeholder Analysis? Everything You Need To Know – Forbes Advisor*. https://www.forbes.com/advisor/business/what-is-stakeholder-analysis/

Hunkenschroer, A. L., & Luetge, C. (2022). Ethics of AI-Enabled Recruiting and Selection: A Review and Research Agenda. *Journal of Business Ethics*, *178*(4), 977–1007. https://doi.org/10.1007/s10551-022-05049-6

Javaid, S. (2024, January 12). *Facial Recognition: Best Practices & Use Cases in 2024*. AIMultiple: High Tech Use Cases & Tools to Grow Your Business. https://research.aimultiple.com/facial-recognition/

Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, *13*(3), 795–848. https://doi.org/10.1007/s40685-020-00134-w

Kumar, V. (2013). *101 Design Methods—A Structured Approach for Driving Innovation in Your Organization*. John Wiley & Sons, Inc. https://learning.oreilly.com/library/view/101-design-methods/9781118083468/cvi.xhtml

Lange, A. R., & Duarte, N. (2018, April 4). Understanding Bias in Algorithmic Design. *ASME ISHOW / IDEA LAB*. https://medium.com/impact-engineered/understanding-bias-in-algorithmic-design-db9847103b6e

Lawton, G. (2022, August 29). *AI hiring bias: Everything you need to know | TechTarget*. HR Software. https://www.techtarget.com/searchhrsoftware/tip/AI-hiring-bias-Everything-you-need-to-know

Lee, N. T., Resnick, P., & Barton, G. (2019, May 22). *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms | Brookings*. https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Ma, Y. (2024). Intelligent Recommendation Method for Talent Resources Based on Big Data. *Intelligent Computing Technology and Automation*. https://doi.org/10.3233/ATDE231176

*Machine Learning and Fairness*. (2021, May 24). Microsoft Research webinars and tutorials. https://youtu.be/ZtN6Qx4KddY?si=kWo9MxkNQ7ko1HUB

Marr, B. (2023, December 12). *AI-Enhanced Employee Onboarding: A New Era In HR Practices*. Forbes. https://www.forbes.com/sites/bernardmarr/2023/12/12/ai-enhanced-employee-onboarding-a-new-era-in-hr-practices/

Naakka, S.-S. (2018). *BIG DATA KILPAILUETUNA SUORAHAKU- KONSULTOINNIN EHDOKASHAUISSA Enemmän, nopeammin ja parempia IT-osaajia big datalla?* [University of Turku]. https://www.utupub.fi/bitstream/handle/10024/145309/Naakka%20Sara-Selia.pdf?sequence=1&isAllowed=y

Ongig. (2024). *Job Description Software | Ongig*. https://www.ongig.com

ResumeMent. (2023, October 16). *From Resumes to Algorithms: The Role of AI in Screening Candidates | LinkedIn*. https://www.linkedin.com/pulse/from-resumes-algorithms-role-ai-screening-candidates-resumement/

Saplicki, C. (2022, November 11). Fairness Explained: Definitions and Metrics. *Medium*. https://medium.com/ibm-data-ai/fairness-explained-definitions-and-metrics-9690f8e0a4ea

SDT. (n.d.). *Tools | Service Design Tools*. Retrieved April 18, 2024, from https://servicedesigntools.org/tools.html

Sheard, N. (2022). Employment Discrimination by Algorithm: Can Anyone Be Held Accountable? *University of New South Wales Law Journal*, *45*(2). https://doi.org/10.53637/XTQY4027

Shipman, M. (2021, March 4). Social media checks can bring bias into hiring. *Futurity*. https://www.futurity.org/cybervetting-human-resources-bias-2527382-2/

Stickdorn, M., Hormess, M. E., Lawrence, A., & Schneider, J. (2018a). *This is Service Design Doing*. O'Reilly Media, Inc. https://learning.oreilly.com/library/view/this-is-service/9781491927175/ch03.html

Stickdorn, M., Hormess, M., Lawrence, A., & Schneider, J. (2018b, July). *#TISDD Method Library*. This Is Service Design Doing. https://www.thisisservicedesigndoing.com/methods

Sukernek, W. (2024). *How to Spot Bias in Hiring*. FloCareer. https://blog.flocareer.com/how-to-spot-bias-in-hiring

Teodorescu, M. (2020). *Fairness Criteria | Exploring Fairness in Machine Learning for International Development | Supplemental Resources*. MIT OpenCourseWare. https://ocw.mit.edu/courses/res-ec-001-exploring-fairness-in-machine-learning-for-international-development-spring-2020/pages/module-three-framework/fairness-criteria/

UNESCO. (2023). *Ethical impact assessment. A tool of the Recommendation on the Ethics of Artificial Intelligence*. UNESCO. https://doi.org/10.54678/YTSA7796

Vivek, R. (2023). Enhancing diversity and reducing bias in recruitment through AI: A review of strategies and challenges. *Информатика. Экономика. Управление - Informatics. Economics. Management*, *2*(4), 0101–0118. https://doi.org/10.47813/2782-5280-2023-2-4-0101-0118

Wallbridge, A. (2023, April 13). *How To Conduct A Bullet-Proof Stakeholder Analysis Matrix In 4 Useful Steps*. TSW Training. https://www.tsw.co.uk/blog/leadership-and-management/stakeholder-analysis-matrix/

Wirtz, D. (2022, August 3). *What Is a Workshop? (+2 Examples) | Facilitator School*. https://www.facilitator.school/blog/what-is-a-workshop

www.recruiter.com. (2023, April 26). *Pros and Cons of Using AI in Recruiting*. Recruiter.Com. https://www.recruiter.com/recruiting/pros-and-cons-of-using-ai-in-recruiting/

YLE. (2023, November 11). *Loimme 13-vuotiaan Ellan – Tiktok tarjosi hänelle sisältöä itsemurhasta ja kalorien laskemisesta*. Yle Uutiset. https://yle.fi/a/74-20059318

Zhong, Z. (2018, October 22). *A Tutorial on Fairness in Machine Learning*. Medium. https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

# Charlie

Co-funded by the European Union

Finansieret af Den Europæiske Union. Synspunkter og meninger, der udtrykkes, er dog kun forfatternes og afspejler ikke nødvendigvis Den Europæiske Unions eller Det Europæiske Forvaltningsorgan for Uddannelse og Kultur (EACEA). Hverken EU eller EACEA kan holdes ansvarlig for dem.

Universitat de les Illes Balears · ISQe · AARHUS UNIVERSITY · VAMK · helixconnect · Charlie

2022-1-ES01-KA220-HED-000085257