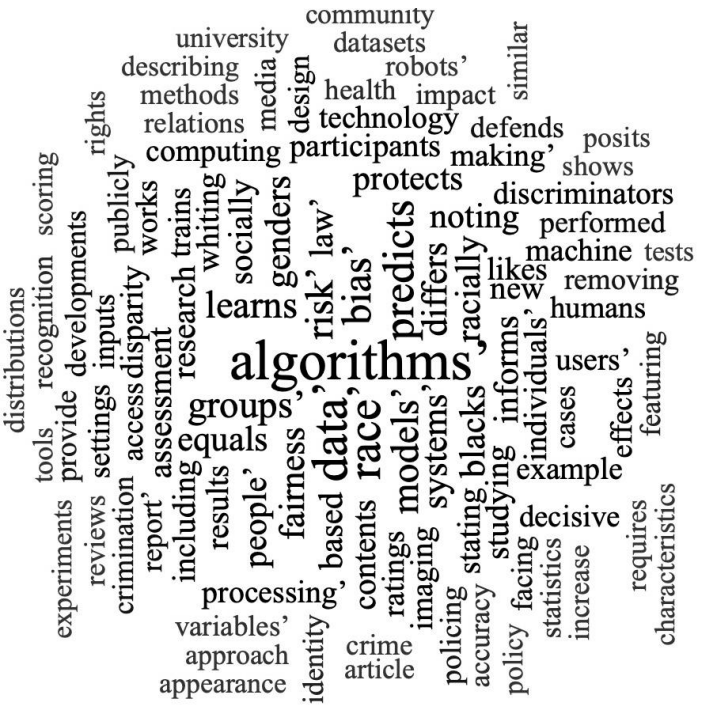


Diversitet ind, diversitet ud

En undersøgelse af bias i design og udvikling af teknologi

SPECIALEOPGAVE

af

Fanny Hvidtfeldt Hesseldahl

Aalborg Universitet

Titelblad

Projektets titel: Diversitet ind, diversitet ud - en undersøgelse af bias i design og udvikling af teknologi

Informationsvidenskab
Specialesemester
Aalborg Universitet, Aalborg

Vejleder: Tanja Svarre

Afleveringsdato: 01/06-2022

Anslag inkl. mellemrum: 186,847
Antal normalsider: 77,8

Abstract

This study addresses biased-related issues in the design and development of technologies such as facial recognition systems, decision-making algorithms, and service robots. The topic is already covered in multiple publications, whereas this research seeks to collect an overview of these issues and combine it with possible solutions through a systematic literature review. It is proven that some of the most widespread technologies used in everyday life and work have better accessibility and functionality for some users rather than others and that these issues are often rooted in biases. The study aims to map out which ethical challenges are characteristic in technology design and development as seen from a user-centered perspective. Ethnicity-related biases and the consequences it entails is the focus of this study. The choice of focus is based on a hypothesis that it seems to be the most critical, as some much-applied technology has shown to function based on properties like facial features including skin tone. Using theories relevant to Human-Centered Design, including Design Thinking and Participatory Design, the study seeks to find answers on how to reduce or even avoid bias-related problems in technology. It also includes an examination of the basic elements of Artificial Intelligence, as this particular technology is an important part of the analysis. The theories are chosen on the basis of the hypothesis that user-centered approaches and values are needed in order to create inclusive designs that are available for all users. Involving users in processes like testing or using representative data training sets should ease the way for technology to work more inclusively. By performing a structured literature review, the study uncovers five main areas of focus presented in the literature, including biases, problem areas, technologies, problem origins, and methods for solving. Each area represents a collection of coded literature relevant to the topic, which creates a categorized overview of each subject. By analyzing this data, four main solution suggestions that can be utilized in developing inclusive ethically considered technology, are discovered: user involvement, representation, transparency, and guidelines. My hope for this thesis is to shed light on the problems currently relevant in technology development and to show that there are established and proven methods available to solve these problems.

Keywords: bias, design thinking, participatory design, human-centered design, technology development, technology design, artificial intelligence, explainable ai

Indholdsfortegnelse

Kapitel 1 - Introduktion……5

1.1 Indledning **5**

1.2 Problemformulering **7**

1.3 Udvidet problemformulering **7**

1.4 Hypotese og motivation for emnet **8**

1.5 Opgavens struktur **9**

Kapitel 2 - Problembehandling……..10

2.1 Metode **10**

2.1.1 Literature Review 10

2.1.2 Litteratursøgning 16

2.1.3 Kodning med NVivo 21

2.1.4 Analyse af materiale 23

2.2 Teori **32**

2.2.1 Kunstig intelligens 32

2.2.2 Human-Centered Design 40

2.2.3 Design Thinking 41

2.2.4 Participatory Design 44

Kapitel 3 - Analyse og resultater……... 47

3.1 Analyse af problemidentificerende materiale **47**

3.1.1 Etiske udfordringer i design og udvikling af teknologi 47

3.1.2 Etniske minoriteter 53

3.2 Analyse af undersøgende materiale **56**

3.2.1 Bias 58

3.2.2 Problemområde 65

3.2.3 Teknologi 67

3.2.4 Problemoprindelse 71

3.2.5 Metode til løsning 76

3.3 Resultater af analyse **83**

3.3.1 Brugerinddragelse 83

3.3.2 Transparens 84

3.3.3 Retningslinjer 85

3.3.4 Repræsentation 86

Kapitel 4 - Diskussion og konklusion……88

4.1 Diskussion **88**

4.2 Konklusion **91**

Denne side er intentionelt blank.

Kapitel 1 - Introduktion

"We replaced all the sensors in the building with a new state-of-the-art system that's gonna save money. It works by detecting light reflected off the skin. It does work, although there is a problem. It doesn't seem to see black people. [...] The company's position is that it's actually the opposite of racist, because it's not targeting black people, it's just ignoring them. [...] they'd like to remind everyone to celebrate the fact that it does see Hispanics, Asians, Pacific Islanders and Jews."

(Better off Ted: Racial Sensitivity, 2009)

1.1 Indledning

Teknologi er kilden til effektivisering, indsigt, tidsbesparing, globalisering - med andre ord, en nemmere hverdag for dens brugere. Udviklingen af teknologiske hjælpemidler har accelererende fart på. Faktisk går det så hurtigt, at det kan være en udfordring for mennesker at følge med. Tendenser som Black Box er blevet mere relevante, hvor maskiner tager beslutninger på et beslutningsgrundlag så kompliceret, at vi ikke kan gennemskue, hvordan den er kommet frem til dens løsninger. Ikke desto mindre anvender vi gerne disse forslag, hvor de bliver placeret i sundhedsvæsenet, politiets arbejdsgang og hjemme i folks stuer.

Teknologiens store gevinster efterlader uundgåeligt spor af udfordringer, intentionelt eller ej. "Hvad vi ikke ser, har vi ikke ondt af" er ikke tilfældet, når algoritmiske beslutninger efterlader marginaliserede målgrupper uden for fællesskabet og mere sårbare, end de var inden teknologiens fremtræden.

Så hvad der der, når teknologi ikke understøtter *alle* brugere? Og hvorfor gør den forskel?

Ovenstående citat er en satirisk repræsentation af reelle problemstillinger, hvor teknologi og dens funktionalitet går hånd i hånd med de biases, den har indlært. Ydre egenskaber som hudfarve og køn har vist at være afgørende faktorer for forskelsbehandling i den måde, teknologi fungerer på. Det er eksempelvis sværere for etniske minoritetsgrupper, særligt kvinder, at anvende basale hverdagssystemer som stemme- og ansigtsgenkendelse. Mere problematisk er det, når deres liv sættes i fare fordi et datasæt undervurderer deres sundhedsbehov. Det er derfor ikke blot et funktionalitetsproblem, men også en udfordring for ligeværdig behandling i et bredere, samfundsmæssigt perspektiv.

I design og udvikling af teknologi er der udarbejdet flere metoder, der baserer sig på brugercentrerede løsninger og værdier. I denne opgave sættes der fokus på Human-Centered Design, Participatory Design og Design Thinking, da disse hører under samme paraply af metoder, der ser brugeren som central deltager i udviklingen af optimale løsninger.

Men hvis der eksisterer metoder med brugercentrerede værdisæt, hvordan kan det så være, at at nogle brugergrupper fortsat forfordeles? Jeg ønsker at undersøge dette paradoks, hvor teknologiudvikling er præget af eksludering af marginaliserede målgrupper på trods af beviser for fordele ved brugercentreret design. Hertil ønsker jeg at afdække hvilke befolkningsgrupper, der er ramt af denne tendens, for dernæst at fokusere på etniske minoritetsgrupper, da denne målgruppe opleves at være undervurderet.

I en informationsvidenskabelig kontekst er det vigtigt at kunne forstå og analysere brugeradfærd- og praksis, hvortil jeg finder det relevant at undersøge de IKT-systemer, hvortil kun nogle målgrupper oplever tilstrækkelig tilgængelighed.

Med specialet håber jeg på at kunne skubbe agendaen om at designe en verden, der er mere tilgængelig for alle.

1.2 Problemformulering

På baggrund af det introducerende interessefelt er følgende problemformulering bestående af tre forskningsspørgsmål udformet:

1. **Hvad kendetegner etiske udfordringer indenfor design og udvikling af teknologi set fra et Human-Centered perspektiv?**

2. **Hvordan gør disse udfordringer sig særligt gældende for brugere på baggrund af etnicitet, og hvad er konsekvenserne?**

3. **Hvordan kan Design Thinking eller lignende processer anvendes til at undgå de belyste problemstillinger?**

1.3 Udvidet problemformulering

Følgende afsnit er en afklaring af hyppigt anvendte begreber samt en afgrænsning af specialets undersøgelse. Afsnittet har til formål at afklare vigtige elementer for læseren.

Begrebsafklaring

- **Bias** anvendes i denne opgave som et generelt term til at beskrive problemer relateret til fordomme på baggrund af demografiske egenskaber om etnicitet og køn. Bias ses i relation til opgavens problemstilling i form af eksempelvis beslutningstagen i AI systemer, der forfordeler en gruppe frem for en anden, særligt i tilfælde hvor det anses som værende uretfærdigt. Med denne definition fokuseres der på hvordan bias indlejres i teknologier og hvad konsekvenser det kan have for brugere.
- **Race og etnicitet**. I 1950 udsendte UNESCO en erklæring, hvori det hævdes at alle mennesker tilhører samme art og at "race" ikke er en biologisk realitet men en myte (Sussman, 2014, s. 1). Selvom denne erklæring kan virke åbenlys i dag, bliver race som term fortsat brugt til at beskrive etnicitets forskelle mellem mennesker. I dette projekt afvises konceptet om race, hvor termet kun anvendes i forbindelse med henvisninger og litteratursøgning, da det ofte indgår i videnskabelige,

engelsksprogede artikler. I opgaven bruges etnicitet i høj grad som betegnelse for den demografiske egenskab, der er central for opgavens undersøgelsesfelt.

- **Udvikling af teknologi.** I brugen af udtrykket "udvikling af teknologi" menes der alt fra programmering til design. Synsvinklen er, at alle der interagerer med en teknologi i udviklingsfasen er medudviklere, da et sådan syn understøtter påstanden om, at alle involverede har et medansvar for det, der bliver udviklet og dertil også for de brugere, det påvirker.

Afgrænsning af emnefelt

Projektopgaven har en brugercentreret vinkel, og tager derfor afstand til tekniske identifikationer som programmering, softwareudvikling og lignende. I analysen er der ligeledes fokus på brugercentrerede teorier og metoder til løsning af de belyste problemstillinger, da det ligger indenfor det pågældende fagområde.

1.4 Hypotese og motivation for emnet

Bias relaterede problemstillinger indenfor design og udvikling af teknologi er en global udfordring i ligestillede samfund. Teknologiens magt og brede anvendelse skaber både nye muligheder og udfordringer, hvor sidstnævnte har vist at være særligt gældende for marginaliserede befolkningsgrupper.

Opgavens motivation udspringer af følgende tre hypoteser:

1) Problemstillinger og følgende konsekvenser indenfor etnisk relateret bias er underbelyst og undervurderet indenfor design og udvikling af teknologi.
2) Designprocesser er præget af ikke at være repræsentativ i enten brugen af data til udvikling af modeller eller i selve brugerinddragelse.
3) Brugercentrerede teorier og principper er nødvendige for at imødekomme problemstillinger relateret til biased teknologi og dens konsekvenser.

Med opgavens problemformulering søges der svar på, om disse hypoteser kan siges at være sande, og i så fald hvad der kan gøres fremadrettet. Mange forskere har allerede søgt dette svar, og disse undersøgelser vil opgaven centrere sig omkring i et systematisk litteraturreview, der kortlægger problemernes eksistens.

Med teknologi har vi muligheden for at vedligeholde, forstærke eller reducere ulighed og diskrimination. Derfor anses det for værende særligt presserende, at undersøge dette fænomen nærmere.

1.5 Opgavens struktur

Følgende kapitel præsenterer opgavens struktur og har til formål at skabe et overblik over projektets indhold.

Nærværende underkapitel er en konklusion på projektets introducerende del, der indleder til problemfeltet. Første kapitel har til formål at indlede læseren i projektets problemfelt og motivation for valg af undersøgelsesfokus. Derudover indeholder den en begrebsafklaring, der skal vejlede læseren i opgavens brug af udvalgte termer.

Andet kapitel indeholder projektets problembehandling, herunder hvilke metoder og teorier der anvendes til analysen.

Tredje kapitel består af analysen af det udvalgte litteratur, der har til formål at besvare projektets problemstilling. Dette gøres med det tidligere præsenteret udvalgt metode og teori.

Fjerde og sidste kapitel indeholder et kritisk perspektiv på projektets resultater herunder en kort diskussion af de mest relevante resultater.

Kapitel 2 - Problembehandling

I dette kapitel præsenteres de metode og teorier, der vil blive anvendt i analysen af det udvalgte litteratur. En overordnet forskningsstrategi bestående af indsamling, fortolkning, analyse og diskussion anvendes således til at tilgå og besvare problemformuleringens forskningsspørgsmål.

2.1 Metode

I følgende afsnit beskrives den metode, der er anvendt for at besvare opgavens problemformulering. Afsnittets længde afspejler den betydning, metoden har haft for specialet, idet det tager udgangspunkt i et systematisk litteraturreview, som beskrevet i nærværende afsnit. Metoden er udvalgt på det grundlag, at et dybdegående og afdækkende litteraturstudie skaber adgang til bred viden og perspektiv, som er nødvendig for at besvare problemformuleringen og nå frem til et tilhørende løsningsforslag.

Afsnittet består af en redegørelse for det systematisk litteraturreview som metode, begrundelse for valg af denne, samt en dokumentation for projektets fremgangsmåde og anvendte søgeprotokol.

2.1.1 Literature Review

Specialets fremstillede problemformulering kræver en sammensætning af både overordnet og specialiseret viden, og for at imødekomme dette er der udført et systematisk litteraturreview. Med denne metode har det været muligt at opnå bred forståelse for emnefeltet og dermed muligheden for at imødekomme udarbejdelsen af et løsningsforslag samt viden til videre undersøgelse.

For at gennemføre litteratur reviewet for nærværende projekt anvendes *The Literature Review* (2012) af Ridley samt *Systematic Approaches to a Successful Literature Review* (2016) af

Booth, Sutton & Papaioannou. Denne viden er blevet brugt til at forberede, organisere og udføre det det litteraturreview, som er grundlaget for opgaven.

For at opnå forståelse for et givent undersøgelsesobjekt, er det vigtigt at udforske det relevante emnefelt, hvori undersøgelsesobjektet ligger. Således bliver det muligt, gennem oplysning, perspektiv og forståelse, at positionere egen forskning i den relevante kontekst. Ved at sætte egen undersøgelse i relation til andres arbejde, kan dette bidrage til merviden, hvilket er formålet med nærværende speciale (Ridley, 2012, s. 1).

Begrundelse for valg af metode

Forskningsemnet centralt for nærværende litteraturreview omhandler en afdækning af hvilke bias, der eksisterer indenfor teknologiudvikling- og design med særlig henblik på radikaliserede minoritetsgrupper. Litteraturreviewet har til formål at afdække dette emne, således at den opstillede problemformulering kan besvares. Det kan understøtte dette formål ved dels at skabe overblik over relateret forskning og dels til at hjælpe med at indsnævre og specificere undersøgelsesområdet. Bias er et kompleks og bredt område, og derfor er det nødvendigt at specificere, særligt grundet projektets begrænsede tid og omfang.

Det udvalgte undersøgelsesområde er bredt og betonet, og projektets undersøgelse eksisterer således ikke isoleret, men i lyset af tidligere forskning og undersøgelser på området. Specialet vil være en tilføjelse til fortsat debat og vidensdeling om emnet (Ridley, 2012, s. 6). Ved at tilføre kontekstuel viden bliver det muligt at se problemet i en større sammenhænge, og dertil nå tættere på en løsning og opnå oplysning.

Litteraturreviewet er dermed et vigtigt element for nærværende opgave, da det kontekstualiserer undersøgelsen, giver overblik over det store landskab af forskning, tilføjer baggrundsviden og dermed skaber plads til denne undersøgelse. Motivationen for opgaven er at bidrage med ny indsigt og perspektiv til allerede eksisterende viden, og derfor har det været særlig vigtigt at udføre et succesfuldt litteraturstudie.

Hvad er et litteraturreview?

Et litteraturreview starter ved første læsning og afsluttes den dag, projektet afsluttes. Det hjælper med at formulere og specificere en given problemstilling undervejs i processen af at

identificere relevante teorier og metoder samt relaterede studier, der kan anvendes til eget studie. Afslutningsvist understøtter det i analysen og fortolkningen af det indsamlede data.

Litteraturreviewet består af to elementer: processen, der er involveret i udførelsen og den endelige, i dette tilfælde, specialerapport. Det involverer relaterede forskningsartikler på undersøgelsesområdet, hvor det forbindes med egen vidensposition. Det udvalgte litteratur anvendes til at identificere teorier og tidligere forskning, og fungerer dermed understøttende for egen identifikation af problemstilling samt pointering af manglende viden eller forskning på det udvalgte område. I specialet er der anvendt publikationer med henblik på at belyse de etiske problemstillinger, der eksisterer indenfor teknologiudvikling, med særligt fokus på racialiserede minoritetsgrupper. Her forbindes tidligere forskning og viden om emnet til egen konklusion og bidrager dermed til nye perspektiver (Ridley, 2012, s. 3).

Litteraturreviewet fungerer således som katalysator for egen undersøgelse, hvormed der opnås ny, samlet viden, der kan anvendes til videre undersøgelse.

Typen af review der er valgt for nærværende opgave, er det systematiske litteraturreview. Systematik er påkrævet i alle typer opgaver af flere årsager. Det hjælper med at skabe klarhed, validere det udførte arbejde, demonstrere stringens i metode, reducerer sandsynligheden for bias og effektivisering af arbejdet (Booth et al., 2016, s. 2, 9). For at sikre effektivitet og høj kvalitet er det nødvendigt med en metode, der tilføjer den rette dybe, sammenhæng, klarhed og effektiv analyse.

Et godt litteraturreview kan resultere i troværdige løsninger på en problemstilling og identificere eventuelle mangler i den eksisterende viden, der kan føre til videre undersøgelse. Ved at identificere nye måder at anskue eksisterende undersøgelser, bliver det muligt at få indsigt i manglende viden på området. Med specialet ønskes at identificere allerede afdækkede emner, så det ikke blot bliver en gentagelse af, hvad der allerede er blevet sagt. Formålet er dermed at resultere i nye perspektiver og indsigter (Booth et al., 2016, s. 11-14).

Booth et al. (2016, s. 19) beskriver tre grunde til at udføre et systematisk litteraturreview:

1. **Klarhed.** Ved at skabe struktur gøres det nemmere at navigere i og fortolke litteraturen. Det hjælper dertil med at indsnævre og klargøre undersøgelseselementet, ved at fokusere på spørgsmål og indrette søgningen derefter. Inklusionskriterier, som

anvendt i denne opgave, gør det nemmere for læseren at forstå, hvorfor noget er til- eller fravalgt. I dette projekt blev der skabt klarhed ved at udarbejde en tabel for anvendte søgetermer opdelt i kategorier, der skabte overblik over søgningens udformning og indhold.

2. **Validering.** En systematisk tilgang kræver, at litteratur udvælges på baggrund af relevans og ikke fordi de har et ønsket udfald eller af personlig interesse. På denne måde undgås potentielle, såkaldt "selection bias" (Booth et al., 2016, s. 19), og opgavens resultater bliver således valideret. Dette er sikret ved udarbejdelse af en søgeprotokol forinden søgningen, anvist i bilag 1, som indeholder dokumentation for søgningerne på tværs af databaser, samt argumentation for eventuelle ændringer. Derudover er der dokumenteret fagligt belæg for denne søgeprotokol, herunder valg af søgetermer og systematik.
3. **Revidering.** Den gennemsigtighed et systematisk review tilbyder, bekræfter at argumentationer ikke er fabrikeret, men at der konkluderes på baggrund af det indhentede data. Dette udføres ved løbende henvisninger til det videnskabelige litteratur, hvor egne argumentationer understøttes af de udvalgte publikationer. Argumentationerne har dermed grobund i litteraturen, og således ikke tilfældig.

I overvejelserne omkring hvilket indhold, der ønskes inkluderet i et litteraturreview, er det vigtigt at reflektere over formålet for litteraturreviewet (Ridley, 2012, s. 23). Ridley beskriver en række formål, hvoraf følgende er udvalgt efter relevans for denne opgave:

- **Overblik.** At skabe overblik over nuværende kontekst for problemstillingen, aktuelle problematikker, spørgsmål og debatter, herunder hvilke emner, der bliver debatteret. Dette er udført ved at lave ordfrekvensanalyse, hvor en illustration af indholdet i hver kategori præsenteres på en overskuelig måde. Dette er samtidig med til at begrunde egne argumenter og giver læseren en fornemmelse af emnets vigtighed (Ridley, 2012, s. 28-30). Under kategorien "Bias" er det eksempelvis tydeliggjort, at etnicitet og køn er de emner, der har mest indhold, hvilket bekræfter projektets problemfelt og fokus.
- **Historisk perspektiv.** At placere problemstillingen i en historisk kontekst, der giver indblik i, hvordan det tidligere har været til hvordan problemet eksisterer i dag. Dette er ikke prioriteret for denne opgave, da det indhentede litteratur har nyere relevans. Dog kan det være til videre undersøgelse, at der i litteratursøgningen blev afgrænset fra årstallene 2000-2022, hvor den ældste anvendte artikel er fra 2011. Her kunne der

være en pointe at undersøge, om emnet ikke er blevet diskuteret eller undersøgt før dette tidspunkt.

- **Diskussion.** At åbne for diskussion af relevante teorier og koncepter, der understøtter egen undersøgelse. I forlængelse defineres teori som et framework, der tilbyder en forklaring, ofte i form af kategorisering samt relation fra et trin til et andet (Ridley, 2012, s. 30-32). I denne opgave anvendes teorier som framework til analyse og fortolkning af de udvalgte artikler, og det åbner for diskussion om hvorvidt de belyste teorier er blevet anvendt i det undersøgende materiale, eller om hvorvidt det med fordel kunne have været anvendt i forskellige tilfælde.
- **Terminologi og definitioner.** At introducere til relevant terminologi og definitioner for at afklare hvilke termer, der bruges i konteksten af undersøgelsen (Ridley, 2012, s. 33-35). Det er vigtigt ikke at antage en fælles forståelse af mening af termer, der anvendes i undersøgelsen, men at læseren introduceres til de variationer af måder, termerne bliver anvendt på i de forskellige artikler samt afklaring af, hvordan egen undersøgelse gør brug af termerne (Ridley, 2012, s 34). Der er tidligere i introduktionen blevet præsenteret for udvalgte termer og deres anvendelse og relevans for nærværende opgave.
- **Identificerer manglende viden.** Ved at beskrive relevant forskning af emnet bliver det påvist, hvordan egen undersøgelse udvider eller udfordrer disse, og dermed adresserer manglende viden eller undersøgelse (Ridley, 2012, s. 35-38). Dertil kan tilføjes en egen motivation for valg af emne, hvilket forefindes i afsnit 1.4.
- **Bevismateriale.** At bevise praktiske problematikker, som undersøgelsen adresserer, og dermed understrege dens relevans (Ridley, 2012, s. 24). Her udgør de indhentede artikler og deres undersøgelser bevis for problemformuleringens relevans.

Disse formål udgør tilsammen grundlaget for, hvorfor det systematiske litteraturreview er valgt som projektets primære metode, da det vurderes, at der med denne kan opnås en troværdig indsigt i problemfeltet. Med et litteraturreview opbygges der en argumentation, der leder til selve undersøgelsen. Denne argumentation udvikles sideløbende med, at relevante kilder inkluderes til at understøtte og udfordre egne påstande. Med argumentationen understreges den tillærte viden om området samtidig med, at der deltages i en dialog med andre forskere om samme emne. Der deltages således i et fagligt fællesskab (Ridley, 2012, s. 24).

Booth et al. præsenterer akronymet "TREAD" (2016, s. 39), der skaber plads til overvejelser, hvilke har relevans for nærværende projektopgaves begrundelse for valg af systematisk litteraturreview. "TREAD" står for følgende:

- **Tid.** Overvejelser om hvor meget tid, der er tilgængelig for udførelse af studiet. Da specialet har en tidshorisont på nogle måneder, har dette indflydelse på hvor lang tid der kan bruges på de enkelte delopgaver i udførelsen af et review.
- **Ressourcer.** Da opgaven udføres af et enkelt medlem, begrænses undersøgelsesområdet til et primært fokus, racialiserede minoritetsgrupper, for at få mest ud af den tilgængelige tid. Det er prioriteret at komme godt omkring et enkelt emne, frem for at overfladebehandle flere emner. Der er her indsnævret fra at gå i dybden med bias mod særligt neurodiverse målgrupper samt brugere med fysiske udfordringer.
- **Ekspertise.** Mangel på tidligere erfaring med udførelse af et så gennemført review har indflydelse på, hvor lang tid de enkelte opgaver kræver. Det har eksempelvis krævet mere tid at få indsigt i hvordan et sådant review gennemføres, end hvis det havde været afprøvet i tidligere opgaver. Denne læringsproces fratager selvsagt tid fra at udføre selve reviewet.
- **Publikum og formål.** I skrivningen er det nødvendigt at overveje hvem modtageren er, i dette tilfælde eksaminator og andre interesser, der forhåbentligt kan få indsigt i emnet (Booth et al., 2016, s. 39). Rapporten kræver nye indsigter samt en redegørelse for allerede eksisterende litteratur, hvilket der stræbes efter. Formålet for denne opgave er særligt at sætte fokus på et allerede afdækket emne, og opdage nye indsigter i de undersøgelser, der findes på området.
- **Data.** Projektet indeholder både kvalitativ og kvantitativ data, hvilket skaber en tyngde og troværdighed. Systematisk review som metode er fleksibelt nok til at kunne indeholde begge typer. Kvantitativ data kræver tal og statistik, hvor kvalitativ kræver en kontekstualisering af den indsamlede data, for at forbedre forståelsen for indholdet (Booth et al., 2016, s. 40). Den kvalitative data er eksempelvis indsamling af dybdegående, eksplorative analyser af brugeres holdning og oplevelse med et system. Kvantitativ data forekommer i form af statistikker, eksempelvis procentdele af hvor godt en teknologi virker for forskellige målgrupper. Begge metoder er en del af det udvalgte litteratur, og skaber tilsammen et beviseligt grundlag for problemstillingerne, de belyser.

2.1.2 Litteratursøgning

I følgende afsnit dokumenteres den udførte søgning for litteraturreviewet for specialerapporten. Søgestrukturen og overvejelserne der udgør denne er dokumenteret. Booth et al. (2016, s. 110) præsenterer nogle grundlæggende trin i en litteratursøgning, der er blevet fulgt:

Trin 1 - Initierende søgning

Først skabes der et overblik over emnefeltets fylde, for at blive bevidst om, hvor meget forskning, der eksisterer på området (Booth et al., 2016, s. 110). Derudover dannes der et overblik over, hvilke søgetermer, der anvendes indenfor fagområdet. Til dette anvendes to generiske database, Scopus og Proquest. Efter denne korte, initierende søgning og overblik afgøres det hvilke databaser, der skal anvendes til den endelige søgning. De udvalgte er vist i tabellen nedenfor.

Kilde	Begrundelse for valg af kilde	Udbyder
Proquest	Generisk, dækker over mange forskellige emner.	Proquest
ACM Digital Library	Dækker teknologi.	ACM
Scopus	Generisk, dækker over mange forskellige emner.	Elsevier

Tabel 1. Udvalgte databaser, deres udbydere samt begrundelse for valg af disse.

Dernæst identificeres relevante nøgleord til anvendelse i søgningen (Booth et al., 2016, s. 110). Disse udvælges på baggrund af problemformuleringen og ud fra et indblik i, hvilke nøgleord relevante artikler anvender fundet i udførte prøvesøgninger. Derudover blev diverse kilder som Google, fagrelevante Podcasts og andet faglitteratur undersøgt for inspiration, og for at få et bedre indblik i, hvilket fagsprog der anvendes på området. Den engelske ordbog blev desuden anvendt til at identificere bøjninger og synonymer, for at undgå at miste relevante søgeresultater på grund af specifikke bøjninger samt undgå blinde vinkler.

De identificerede nøgleord blev sorteret i fire aspekter, der hver indeholder relevante termer. Dette for at skabe struktur, overblik og mulighed for fleksibilitet i selve søgningen.

Eventuelle forkortelser af de opstillede termer er anvendt i den endelige søgning, men ikke vist udtømmende i tabel 2.

Aspekt 1 Fagområde	Aspekt 2 Målgruppe	Aspekt 3 Problem	Aspekt 4 Løsning
Human Computer Interaction UX Design User Experience Artificial Intelligence Algorithms Technology Predictive Software	Minorities Neurodiversity Ethnic Gender Race	Bias Exclusion Discrimination Inequality Oppression Availability Bias Status Quo Bias	Accessibility Human Rights Equality Inclusion Design For All Universal Design

Tabel 2: Anvendte søgetermer opdelt i aspekter

Efter at have etableret databaser og søgetermer udvikledes en søgestrategi (Booth et al., 2016, s. 110). I bilag 1 ses en tabel over de udførte søgninger, herunder hvilke søgestrategier der blev udført i hvilke databaser, antal resultater, dato for søgning, relevante detaljer samt andre bemærkninger. Den samme søgning er blevet anvendt over alle tre databaser med få undtagelser. Ved databaserne Proquest og Scopus bemærkedes det at aspekt 2, målgruppe, blev begrænset til det udvalgte bias for at opnå færre, mere relevante resultater. Dette grundet databasernes generiske indhold. I databasen ACM blev aspekt 1, fagområde, fjernet fra søgningen, da databasen har et teknologisk fokus, og det deraf gav færre, mere relevante resultater. Aspektet fagområde indeholder termer der understøtter egen faglighed samt problemstillingen. Termerne repræsenterer et fagligt relevant perspektiv, men har også termet "technology" med, der kan gå ud over egen faglighed. Dette er for at skabe relevans i databaser som Proquest og Scopus, der har fagligt bredt indhold.

I bilag 1 ses også anvendelsen af avanceret søgning i databaserne. Her anvendtes "AND" til at adskille aspekterne fra hinanden og søgningen blev begrænset til, at søgetermerne skulle indgå i publikationernes abstracts. Specielle karakterer, * og ", anvendtes for at inkludere bøjninger af ord samt for at sammensætte flere ord.

Trin 2 - Endelige søgning

Andet trin består af selve søgningen. Her søges i alle udvalgte databaser ved brug af den udarbejdede søgestrategi. Undervejs dokumenteres eventuelle ændringer i søgningen, anvendelse af filtrering mm., som anført i forrige trin. Søgningerne gav i alt 473 resultater fordelt over de tre databaser.

Trin 3 - Bibliografisk søgning

Tredje trin består af en søgning ud fra referencelister fundet i det udvalgte litteratur. Idet der allerede var fundet en stor mængde litteratur, relativt til opgavens tidsramme og at der kun er én person på projektet, blev dette trin udført overfladisk, herunder en hurtig gennemgang af de mest relevante publikationer og indeholdende referencelister, hvoraf der ikke blev tilføjet til den endelige litteraturliste. Søgningen tilføjede dog indsigt til det allerede funden litteratur, og bekræftede på den måde det udvalgte litteraturs relevans.

Trin 4 - Verifikation

I fjerde trin verificeres søgestrategien, og der undersøges om relevante artikler, der ikke fandtes under søgestrategien, kan være relevante at inddrage. Det var ikke tilfældet.

Trin 5 - Dokumentation

I femte trin dokumenteres søgestrategien, som illustreret løbende i dette kapitel (Booth et al., 2016, s. 110).

Efter søgning og dokumentation følger en udvælgelsesproces, der skal sikre inkludering af relevante, udtømmende litteratur indenfor emnet. Til dette er et reference styringssoftware, Zotero, anvendt, da det fungerer både administrerende, kategoriserende samt understøtter korrekt litteraturhenvisning.

Forinden udvælgelsen blev der udarbejdet inklusionskriterier (Booth et al., 2016, s. 85), for at afdække forskningsemnet korrekt. Kriterierne inkluderer følgende:

- Sproget skal være dansk eller engelsk.
- Udgivelsesåret skal være indenfor 2000-2022 af hensyn til relevans for emnet. Fra omkring år 2000 begyndte internettet at blive alment og computerne fik plads i private hjem. I forlængelse med dette blev sociale medier også mere udbredt.

- Publikationerne skal være Peer Reviewed. I tilfælde hvor det ikke var muligt at filtrere efter dette kriterium i en database, blev artiklerne efterfølgende undersøgt manuelt for kriteriet og sorteret derefter.
- Termer indenfor aspekterne teknologi og bias skal indgå (se tabel 2).
- Har relevans indenfor opgavens teoretiske ramme. Dette kan komme til udtryk på flere måder: at et eller flere elementer fra teorierne bevidst bliver anvendt empirisk eller nævnt på anden vis; at publikationens indhold kan benyttes til at informere om teorien, og således vise hvordan teoriens principper er relevante; mulighed for analyse af hvorledes teorien kunne have været anvendt i den givne undersøgelse.

Udvælgelsen blev udført af tre omgange (Booth et. al, 2016, s. 143). Først ud fra titel, hvor artikler blev fravalgt ud fra vurdering af titlens relevans og de nævnte inklusionskriterier. Da der er risiko ved at determinere indhold på baggrund af titel, blev der lavet en mild frasortering, hvoraf kun de der uden tvivl ikke levede op til kriterierne på baggrund af titlen, blev fravalgt. Dette for at undgå frasortering af ellers relevante publikationer. Hvis der opstod tvivl i første sortering, blev disse altså gemt til anden sorteringsrunde. Efter første sortering var der i alt 143 publikationer tilbage, heraf 28 fra Proquest, 51 fra ACM og 64 fra Scopus.

Anden sortering blev udført på samme måde som ovenstående, blot ud fra artiklernes abstract. Efter denne sortering var der i alt 42 publikationer tilbage, heraf 16 fra Proquest, 14 fra ACM og 12 fra Scopus.

Tredje og endelige sortering blev udført ud fra det fulde indhold. Publikationerne blev gennemgået ud fra hele teksten, og der dokumenteres begrundelse for fravælgelse. Under denne sortering blev i alt 9 artikler fravalgt. Resultatet blev dermed 33 antal publikationer ud af de 473 indhentede. Begrundelsen for frasortering er baseret ud fra inklusionskriterierne samt løbende dokumenterede fravælgelseskriterier på baggrund af artiklernes emne og hovedfokus. Når de nedenstående fravalgte emner har være hovedfokusset i artikler, er de blevet frasorteret, da de er vurderet for irrelevante i forhold til projektets emne. De frasorterede emner er følgende:

- Ligeløn.
- Integration.
- Miljø og klima.
- Historisk fokus.

- Sport og/eller genetik.
- Politik og/eller procedurer.
- Specifikt geografisk fokus.
- Uddannelse og/eller klassemiljø.
- Covid-19 og online undervisning.
- Artikler med fokus på softwareudvikling.
- Arbejdskraft og/eller relaterede magtrelationer.

Derudover blev to artikler fravalgt grundet manglende mulighed for at indhente den fulde tekst.

Det endelige udvalgte litteratur blev efterfølgende opdelt i to grupper, da disse to grupper repræsenterer litteratur med hver deres formål. Dette for at skabe et overblik over artiklernes indhold og brugbarhed i relation til at besvare problemformuleringen. Grupperne blev opdelt i kategorierne *Problemidentificerende* og *Undersøgende*.

Kategorien **problemidentificerende** indeholder artikler af beskrivende karakter, der anvendes til en redegørelse for hvilke bias udfordringer, der eksisterer indenfor udvikling og design af teknologi. Disse anvendes primært til at besvare det første spørgsmål i problemformuleringen, som består af en redegørelse og afdækning af bias i teknologiudvikling.

Kategorien **undersøgende** er artikler der i højere grad end de problemidentificerende indeholder konkrete, empiriske undersøgelser, og som kan anvendes til en vurdering af, hvorvidt Design Thinking eller lignende processer relaterer sig til empirien. Derudover til en diskussion af hvorvidt Design Thinking eller andre teorier kunne have været anvendt for bedre resultater. De undersøgende artikler anvendes primært til at besvare problemformuleringens sidste spørgsmål, som er løsningsfokuseret. Disse artikler tilbyder konkrete undersøgelser og resultater samt handlingsforslag, hvorledes det er muligt at se på relationen mellem disse og belyste designteorier.

14 artikler blev tildelt den problemidentificerende gruppe, og de resterende 19 tildelt den undersøgende.

Begge grupper af materiale blev behandlet og analyseret ved brug af analysemetoden præsenteret senere i opgaven efter udviklede koderammer (afsnit 2.1.4).

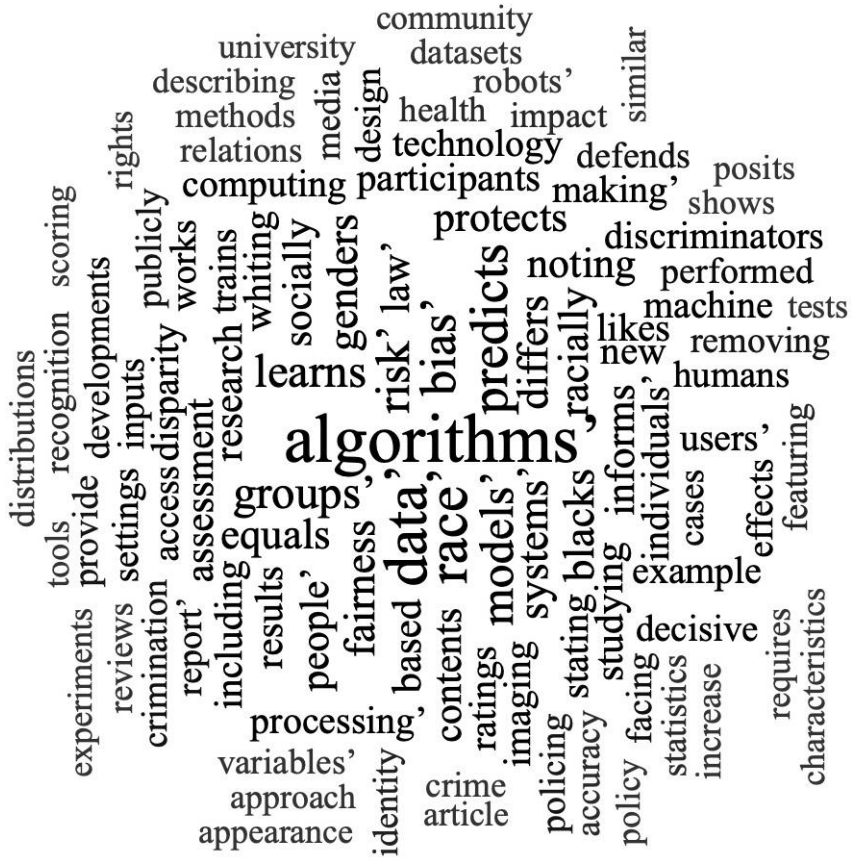
2.1.3 Kodning med NVivo

I analyseprocessen af det indhentede litteratur er programmet NVivo anvendt, som er et kvalitativt dataanalyse software program. Programmet kan understøtte analyseprocesser ved at skabe dybere indsigter, øge produktivitet og fremme konklusioner. Programmet kan indhente data, i denne opgave litteraturen, fra forskellige kilder, i dette tilfælde Zotero, og tilbyder forskellige analysefunktioner som styrings,- noterings,- og visualiseringsværktøjer.

Programmet er anvendt da det har flere fordele. Det kan det skabe et kvantitativt overblik over hvilke emner og temaer, der forekommer mest i de indhentede publikationer, og har samtidig gode, kvalitative analysemuligheder, hvilket er primært for opgaven.

Som en indledning til analyse af det indhentede materiale, blev programmet anvendt til at lave en såkaldt "word frequency analysis", hvor programmet analyserer alle artikler og udvælger de 100 mest forekommende ord. Analysen inkluderede bøjninger af ordene, havde en begrænsning på minimum længde af 3 bogstaver, og der blev anvendt en "stop word list", som er en liste over ord, der ekskluderes fra analysen (heriblandt årstal og konjunktioner).

Resultatet fra analysen af de mest forekommende ord kan ses i billede 1 samt som liste i bilag 2. Det skaber et overblik over de hyppigste temaer og emner, som de indhentede publikationer indeholder. Her var ordene algorithms, data, race, predicts, og bias de fem mest forekommende ord, hvor alle fremkom mere end tusinde gange (inklusiv bøjninger). Det bekræfter, at artiklernes indhold stemmer overens med det, opgaven ønsker at undersøge - algoritmiske forudsigelser og data, der har bias relaterede problematiske udfald på baggrund af bias. Hvor etnicitet er fokuspunktet i opgaven, viser dette sig også i denne ordanalyse, hvor relaterede ord som "racially" er nogle af de mest hyppigt forekommende ord i publikationerne.



Billede 1. Resultat af ordfrekvensanalyse

I behandling af det problemidentificerende materiale blev en ordfrekvensanalyse også udført. Denne viste, at kunstig intelligens, efterfølgende AI, og tilhørende paraplybegreber eller synonymer som "AI", "algoritmer", "data", "machine learning" og "predictive" er nogle af de mest forekommende ord i de 11 artikler behandlet under emnet *Etiske udfordringer i design og udvikling af teknologi*. Analysen placerer ordet "algorithms" i toppen med i alt 76 forekomster. Dertil forekommer begreberne "ai" og "artificial intelligence" tilsammen 78 gange. Det bekræfter artiklernes relevans for problemstillingen og undersøgelsesområdet. AI er et vidt begreb der dækker bredt og anvendes i forskellige sammenhænge i forskellige teknologiske løsninger, og derfor er det også forudsigeligt, at begrebet skulle forekomme i høj grad.

Ordfrekvensen er således anvendt til at skabe et hurtigt overbliksbillede over indholdet i artiklerne samt bekræfte deres relevans.

Programmet er således anvendt til at understøtte analyse samt gennemarbejde af artiklerne, da det er muligt at, som før nævnt, kode artiklerne og dermed linke beviser fra en artikel til en

anden. Der kan tilføjes annotationer, hvor notater tilføjes direkte til en sætning eller enkelte ord, overstreges direkte i artiklerne samt andre funktioner, der hjælper processen med at analysere, fortolke og sammenligne artiklerne. Derudover kan indholdet inddeles og kategoriseres i mapper, hvor de i dette tilfælde er inddelt i problemidentificerende og undersøgende artikler, som tidligere beskrevet.

2.1.4 Analyse af materiale

Til analyse af de udvalgte publikationer er Schreiers (2014) beskrivelse af kvalitativ indholdsanalyse anvendt, da den tilbyder en struktureret tilgang til behandling og kodning af det udvalgte materiale. Der vil i følgende afsnit blive gennemgået de punkter, analysen lægger op til, og dertil illustreres hvorledes metoden er blevet anvendt i denne opgave.

Den kvalitative indholdsanalyse er en systematisk metode, der beskriver meningen af kvalitativ data (Schreier, 2014, s. 2). Denne opgaves grundlæggende metode er det strukturerede litteraturreview, som er systematisk og gennemsigtigt, hvorledes den kvalitative indholdsanalyse understøtter disse værdier. Strukturen opstår i at kategorisere det udvalgte materiale, litteraturen, i en kodningsramme. Strukturen karakteriseres af tre egenskaber:

Reducering af data. Den kvalitative indholdsanalyse reducerer mængden af data, og skaber dermed mulighed for at fokusere på udvalgte aspekter, der er rettet mod problemstillingen. Mængden af aspekter, der undersøges, reduceres til et antal, der er muligt for forskeren at håndtere, hvilket har været optimalt for nærværende opgave, som udarbejdes af en person. Reduceringen opstår i kategoriseringen, hvor disse beskrives overordnet og abstrakt, og dermed kan indeholde flere forskellige passager, frem for mange konkrete kategorier (Schreier, s. 2). I dette tilfælde blev kategoriseringen begrænset til maksimalt seks underkategorier i hver af de udarbejdede fem hovedkategorier.

Systematik. Al materiale gennemarbejdes og undersøges for, hvilket indhold der er relevant for problemstillingen. Ved at følge en systematisk proces med specifikke trin, undgås muligheden for at undersøge materialet gennem egne antagelser og forventninger. Processen kan være iterativ, hvilket den var for nærværende projekt, hvor kodningsrammen blev ændret undervejs, som materialet blev behandlet. Selvom der vælges en iterativ tilgang, er trinene, der skal udføres de samme, hvilket bekræfter den systematiske egenskab. Systematikken

opstår blandt andet i kodningen, som gennemføres to gange for at kvalitetssikre definitionerne af kategorierne, som skal være klare og entydige (Schreier, 2014, s. 2-3). Kodningen blev i dette tilfælde udført undervejs, hvor kvalitetssikringen primært foregik i revideringstrinnet, som belyses senere i dette afsnit. Der kodes således ikke dobbelt, men i stedet revideres koderne efter al materiale er behandlet.

Fleksibilitet. Ved at kombinere koncept- og datadrevne kategorier i samme kodningsramme, skabes der fleksibilitet i forhold til at sikre, at kodningsrammen altid matcher materialet og ikke omvendt. Kategorierne vil dermed indeholde dele af begge, og samtidig sikre at matche dataene (Schreier, 2014, s. 3). Rammen der er opstillet til at kode materialet fungerer således som et udgangspunkt, men udelukker ikke eventuelle andre fund, der kunne være relevante for problemstillingen. Der bliver dermed åbnet op for andre relevante aspekter, der går udenom det oprindelige teoretiske standpunkt, således at dataene får tilladelse til at fortælle noget om materialet, som går udenom det teoretiske fundament. Dermed er der ikke noget materiale, der i forvejen udelukkes, og det sikres at alt relevant materiale medtages.

I følgende beskrives hvorledes den kvalitative indholdsanalyse udføres ifølge Schreier (2014), samt hvordan metoden er anvendt for denne opgave.

Analysen foregår i alt i otte trin:
1. Udarbejdelse af forskningsspørgsmål. Dette er gjort forinden, som består af den udarbejdede problemformulering. Det er disse spørgsmål, analysen søger at besvare.

2. Udvælgelse af materiale. Dette er gjort ved den tidligere beskrevet litteratursøgning og dertil søgeprotokol.

3. Opbygning af koderamme. I dette trin udarbejdes den kodningsramme, der anvendes i analysen af materialet. Koderammen udgør, hvordan selve kodningen udføres i analysen, og beskrives som værende central i metoden for den kvalitative indholdsanalyse.

En kodning består af minimum én hovedkategori, gerne flere, samt minimum to underkategorier, som udspringer af førstnævnte. En hovedkategori afdækker et enkelt aspekt fra det udvalgte materiale, og indeholder dermed materiale, der omhandler undersøgelsesfokusset (Schreier, 2014, s. 7).

Til opgavens analyse er der udarbejdet i alt fem hovedkategorier navngivet Bias, Problemområde, Teknologi, Problem Oprindelse og Metode, som består af materiale, der tilhører disse kategorier. Kategorierne er tildelt en beskrivelse af, hvilke kriterier der kræves, for at en passage kan tilhøre denne. Dette er illustreret i den fulde koderamme, anvist i bilag 3. Hovedkategorierne indeholder ikke i sig selv koder, men fungerer som et samlingspunkt for de senere beskrevne underkategorier, og er med til at danne struktur og overblik over kodningerne.

De udarbejdede hovedkategorier er følgende:
Bias indeholder passager, der beskriver hvilket bias, der bliver omtalt samt dennes konsekvenser. Kategorien indeholder således konkrete eksempler på konsekvenser af den pågældende bias, samt generel oplysning om denne. Derudover indeholder den også generel information om begrebet bias og dets betydning. Kategorien er med til at skabe overblik over, hvor mange forskellige bias der bliver nævnt og i hvilket omfang, og har dermed en kvantitativ funktion, der kan analyseres ud fra. Den er desuden afgørende for at besvare problemformuleringens første, afdækkende spørgsmål, som søger at afdække hvilke etiske udfordringer, der eksisterer indenfor design af teknologi.

Problemområde har til formål at skabe overblik over i hvilke samfundsmæssige områder, udfordringerne opstår. Det inkluderer passager, der beskriver hvilket område, den givne problematik er gældende i, eksempelvis sundhedssektoren. Det beskriver således både hvor problemerne opstår henne, samt hvordan den beskrevne teknologi bliver anvendt. Det kan særligt være med til at besvare problemformuleringens andet spørgsmål, der søger svar på hvordan udfordringerne gør sig gældende for etniske minoriteter, og giver dertil også et kvantitativt billede af, hvor problemerne opstår.

Teknologi som kategori er problemorienteret, og beskriver hvilken teknologi der beskrives og hvilke problemer der er med denne. Det skaber overblik over, hvilke teknologier der bliver belyst i artiklerne, og som der er udfordringer med i forbindelse med anvendelse eller lignende. Et eksempel på dette er AI, hvor forskellige artikler beskriver teknologiens potentiale til enten at opretholde eller reducere systematiske uligheder.

Problemoprindelse er en kollektion af passager, der beskriver hvorfor, hvornår og/eller hvordan det pågældende problem opstår. Kategorien kom til senere i processen, og var altså ikke oprindeligt oprettet forud for kodningsprocessen. Den opstod grundet et løbende behov for at adskille koder, der beskrev selve problemet fra beskrivelser af hvorfor disse problemer opstår. Viden om hvorfor et problem opstår, er vigtigt for at kunne nå frem til en løsning, hvilket er et vigtigt element i denne opgave.

Metode som kategori er modsat teknologi løsningsorienteret. Den beskriver hvilke metoder eller undersøgelser, der er anvendt for at løse den beskrevne udfordring. Den har til formål at vejlede problemformuleringens sidste spørgsmål, der omhandler hvordan Design Thinking kan anvendes for at løse de belyste udfordringer. Formålet med kategorien var, at finde ud af hvordan denne specifikke teori blev anvendt, men det blev opdaget, at denne ikke nævnes eksplicit i artiklerne. Derfor fungerer den oplysende og vejledende, hvor den oplyser om anvendte metoder til løsning og samtidig vejleder i refleksion over, hvordan Design Thinking kan spille en rolle i løsninger.

Underkategorier specificerer nærmere, hvad materialet indeholder, og har relation til den tilhørende hovedkategori (Schreier, 2014, s. 7). Hertil er der udarbejdet 21 underkategorier fordelt på hovedkategorierne, som er nærmere konkretiseret. Et eksempel er underkategorier er under Bias, der indeholder underkategorierne Etnicitet, Fysisk handicap, Køn, samt Generelt om bias. På denne måde er det muligt at have alle belyste bias elementer samlet ét sted, fordelt ud over konkrete emner.

Sidst er der krav om, at alle relevante aspekter af materialet skal være dækket af en kategori for at sikre at alt materiale er ligeligt medregnet. Det er muligt at lave såkaldte restkategorier, hvis der er dele af materialet, der ikke kan placeres i andre kategorier, for at efterleve dette krav. Disse skal dog bruges sparsomt (Schreier, 2014, s. 8). I dette tilfælde blev der oprindeligt lavet en restkategori, hvor denne bruges ved tvivl om, hvor en relevant passage hørte til. Denne blev dog gennemgået som afslutning på kodningsprocessen, hvor alle kategorier var udarbejdet, testet og anvendt, og der dermed var større sikkerhed i at anvende disse. I gennemgangen var det muligt at placere alle passager fra restkategorien til relevante kategorier, og den blev dermed udtømt for indhold og fjernet.

Kodningsrammen for nærværende analyse består dermed af to hierakier (en hoved- og underkategori), indeholdende i alt fem hovedkategorier og 21 underkategorier fordelt nogenlunde ligeligt på disse. Hver hovedkategori indeholder mellem fire og fem underkategorier.

Den viste koderamme er udviklet ved at følge fire trin. Først en udvælgelse af materiale, hvor der ofte kun anvendes en del af materialet til denne opbygning, for at undgå kognitiv overbelastning (Schreier, 2014, s. 8). Ved at bruge fire artikler var det muligt at udvikle en koderamme, der afdækkede de nødvendige elementer. Der nåede således et mætningspunkt for udbygningen af koderammen ved brug af udvalgte artikler, hvorledes der derfor ikke behøves at anvende det fulde materiale til dette formål. De udvalgte artikler blev valgt på baggrund af at de havde forskelligt, fyldestgørende indhold, som dermed kunne afdække flere kategorier.

Dernæst skabes der strukturering gennem opbygningen af hovedkategorier samt generering gennem opbygning af underkategorier. Dette trin kan udføres på to måder, konceptdreven og datadreven, hvor denne opgave vil følge en blanding af disse to, for at sikre al materiale bliver involveret og beskrevet fyldestgørende (Schreier, 2014, s. 9). Den konceptdrevne tilgang baserer kategorierne på forudgående viden, og bliver i dette tilfælde anvendt til initierende at etablere hovedkategorierne, som udspringer af viden fra litteratursøgningen samt forudgående undersøgelse i form af læsning af relevant materiale for problemstillingen. Inspiration til disse udsprang blandt andet af den ikke videnskabelige, dog evidensbaserede bog *Ruined by Design* (2019), der belyser en bred vifte af uetisk designudvikling og efterfølgende konsekvenser, samt forslag til hvordan disse udfordringer kan løses. Derudover fungerede selve litteratursøgningen som inspiration til, hvilke emner og indhold der kunne være relevant som kategorier. Kategorierne udspringer således af viden, der er forudgående for komplet læsning og bearbejdelse af materialets fulde indhold, men som vurderes være de mest væsentlige overemner.

Den datadrevne tilgang kan anvendes med forskellige strategier, hvoraf der i denne opgave er anvendt en struktureret tilgang præsenteret af Schreier (2014, s. 9). Denne har relevans for nærværende analyse, da den blandt andet giver mulighed for at skabe overordnede kategorier. Strategien er brugbar til datadreven udvikling af underkategorierne, og understøtter dermed den konceptdrevne tilgang, som blev anvendt til hovedkategorierne. En datadreven tilgang er

vigtig for nærværende litteraturreview, da det er det udvalgte materiale, der vægter højest i besvarelsen af problemstillingen. Strategien involverer en undersøgelse af passager, hvor følgende udføres i nævnt rækkefølge:

1. Læsning af materialet indtil relevant koncept opdages.
2. Undersøg om der er en underkategori, der passer til dette koncept.
3. Hvis der er lavet en underkategori der passer, tilføjes denne dertil.
4. Hvis ikke, skal der oprettes en ny underkategori der dækker konceptet.
5. Læsningen fortsætter indtil næste relevant koncept opdages.

Strategien udføres indtil der ikke er flere nye koncepter, at opdage, og materialet dermed kan konkluderes at være udtømt for relevant indhold (Schreier, 2014, s. 9). Til at udføre strategien anvendes som tidligere beskrevet programmet NVivo, der gør det nemt og overskueligt at se oprettede kategorier og tildele passager hertil. Dette gøres undervejs i læsningen og udgør dermed måden hvorpå analysen gennemføres.

Tredje trin i kodningsrammens opbygning er en definering af kategorierne, både hoved- og underkategorier, som består af følgende:

- Kategoriens navn, som skal være en kortfattet beskrivelse af, hvad kategorien refererer til.
- Beskrivelse af, hvad der menes med navnet. Her er en definition obligatorisk, og skal beskrive, hvad der menes med den givne kategori, herunder hvilke egenskaber, der er karakteriserende. Dette er illustreret i bilag 3, kodningsrammen.
- Eksempler, der understøtter illustrationen af kategoriens definitioner. Eksemplerne skal være fra eget materiale.
- Sidst kan der valgfrit indskrives beslutningsregler, som specificerer hvad der ikke skal inkluderes samt hvilken kategori, der skal bruges i stedet. Det kan være en hjælp i at afgøre, om en passage passer ind i en given kategori. Dette sikrer at underkategorierne er korrekt tilpasser hovedkategorierne (Schreier, 2014, s. 10-11). Hertil blev inklusionskriterier anvendt for at sikre relevans i det udvalgte materiale.

Det fjerde og sidste trin i udviklingen af kodningsrammen består af en revidering og udvidelse. Trinnet er afgørende for at sikre, at kodningsrammen er korrekt og relevant udført, og dermed er klar til at blive anvendt i analysen. Efter kategorierne er genereret og defineret, ses der med nye øjne på strukturen, og der ryddes op i eventuelle løse ender. Hvis to

underkategorier er meget ens, kan det give mening at sammensætte disse. Det kan også være, at en underkategori er mere omfattende end andre, dertil kan denne gøres til en hovedkategori (Schreier, 2014, s. 11). På denne måde revideres den opbyggede struktur. I dette tilfælde blev alle kategorierne som nævnt gennemgået efter alt materiale var kodet, hvor restkategorien blev udtømt og fjernet. Koderne blev derudover gennemgået overordnet for indhold og indholdets placering. Her blev det blandt andet opdaget, at en af underkategorierne kun indeholdte én enkelt passage. Denne passage viste sig at være relevant i en anden kategori, og den pågældende underkategori kunne dermed fjernes. At en underkategori kun indeholder én passage er et tegn på, at den ikke er relevant, og kan indgå et andet sted. Derudover blev der fjernet i alt seks underkategorier, da disse indeholdte få koder der havde gentagen karakter. Det vurderes derfor, at disse ville være fyld og ikke tilføje merviden til opgaven. På den måde fungerede NVivo på flere måder som et understøttende værktøj, der skabte et hurtigt og overskueligt overblik over alle kategorierne.

Udvidelse består i at indvie resterende data, der indtil nu ikke har været inkluderet. Dette gøres ved en gennemgang tidligere trin, for at sikre, at alt data er inkluderet. Det var ikke nødvendigt i dette tilfælde, da alt materiale blev dækket på én gang.

Kodningsramme for det problemidentificerende materiale

Det bemærkes, at den beskrevne opbygning af kodningsramme er blevet anvendt for det undersøgende materiale. Til det problemidentificerende materiale er en simplificeret udgave anvendt, da dette materiale bruges redegørende og det derfor anses for værende tilfredsstillende, at anvende en forkortet version. I behandling af det problemidentificerende blev litteraturen kodet i to overordnede kategorier baseret på opgavens problemformulering. Kategorierne repræsenterer litteraturens indhold opdelt i relevante emner, således at det skaber en struktur i beskrivelsen heraf. Det bemærkes desuden, at problemidentificerende litteratur i nogle tilfælde blev tildelt koder fra den undersøgende proces, som var de primære koder. Dette for ikke at udelukke noget materiale, da elementer i det problemidentificerende materiale viste sig at være relevant for den primære analyse.

4. og 5. og 6. fase - segmentering, prøvekodning og evaluering. Fjerde, femte og sjette fase i analysemetoden er kun lettere berørte, da disse er mest velegnede til store datasæt og en bredere tidsramme, og derfor også slået sammen i deres beskrivelse. Det er vurderet, at den udarbejdede kodningsramme er kvalitetssikret i form af revideringstrinnet, hvor rammen blev

gennemgået og rettet. Derudover er metoden udført lidt anderledes end beskrevet, hvor kodningsrammen er udviklet løbende, således at den kontinuerligt er blevet vurderet og tilrettelagt.

Segmenteringsfasen omfatter to runder af kodninger, hvor disse, såfremt der blot er én forsker på projektet hvilket er tilfældet, foregår på to forskellige tidspunkter. Det bliver dermed muligt at afgøre, om kodningerne er korrekt tildelte. Deraf segmentering, da materialet skal opdeles i enheder for at udføre dette, hvor hver del passer i en specifik underkategori (Schreier, 2014, s. 11-12). Da kodningerne har en overskuelig størrelsesorden, er dette ikke anset for værende nødvendigt, da det løbende har været muligt at udføre kvalitetstjek af korrekte angivelser.

Prøvekodning omfatter en prøveanalyse af dele af materialet. Dette er vigtigt for at kunne opdage eventuelle fejl og rette disse, inden den faktiske analyse (Schreier, 2014, s. 12). Materialet, der udvælges, skal afdække en bred type af data og datakilder i materialet, og vælges ud fra at kunne tilpasses til størstedelen af kategorierne i kodningsrammen, så disse kan tildeles løbende i prøvekodningen. Hertil blev fire artikler udvalgt på baggrund af et kriterie om at indeholder både undersøgende og problemidentificerende materiale. Derudover blev de valgt ud fra en vurdering af, at de indeholdte forskellige indhold i form af den teknologi, der blev belyst, og de metoder, der blev anvendt. Hensigten var, at udvælge materiale der under prøvekodningen kunne opfange og indramme så mange kodningskategorier, som muligt. Dermed var det også muligt at prøvekode flere kategorier og underkategorier undervejs. Processen gav samtidig et indblik i og sikkerhed med at bruge programmet NVivo. Det lykkedes ud fra de fire artikler at udvikle hele kodningsrammen.

Kategorierne fra koderammen blev tilføjet til materialet i løbet af to runder kodning, således at korrekt angivelse sikres (Schreier, 2014, s. 13). Ifølge metoden tager denne proces 10 til 14 dage, hvilket også begrunder den forenklede version anvendt til dette projekt, der bruge en enkelt dag på prøvekodningen. Denne gav dog en del indsigt i hvilke kategorier og underkategorier, der var relevante at udvikle, og hjalp med at forstå hvordan kodningssystemet fungerer i NVivo, hvilket skabte en sikkerhed og tryghed i arbejdet med programmet.

Hvor en af de forud oprettede hovedkategorier indeholdte i alt fire hierarkier, blev det opdaget gennem prøvekodningen, at dette ikke var nødvendigt og faktisk gjorde processen og kodningsstrukturen mere uoverskuelig. Denne blev derfor ændret til en mere relevant opdeling.

Evaluering og modificering af koderammen er en del af prøvekodningen, hvor resultaterne af prøvekodningen efterses for gyldighed og om der er overensstemmelse i kodningerne (Schreier, 2014, s. 13). Hertil blev det vurderet, at kodningen var korrekt udført, hvilket bekræfter kodningsrammen opsætning og dermed dens kvalitet. Derudover sikres validiteten ved at kategorierne er tilstrækkeligt beskrevet, da det erfares hvorvidt de er intuitive at anvende (Schreier, 2014, s. 13-14).

7. Hovedanalyse. Når kodningsrammen er opsat og vurderet klar til anvendelse, skal alt materiale analyseres. Herfra kan kodningsrammen, ifølge metoden, ikke længere ændres, og dertil er kvalitetssikringen vigtig (Schreier, 2014, s. 14). Da prøvekodningen i denne opgave var nedprioriteteret, er der valgt en fleksibel tilgang i forhold til at tage en iterativ tilgang til kodningrammens udvikling. Materialet tildeles løbende kategorier og underkategorier, og resultaterne tilføjes til et kodningsskema vist i bilag 3 (Schreier, 2014, s. 14). Under kodningen kan samme artikel fremgå i flere forskellige kategorier.

8. Præsentation og fortolkning af fund. Kodningsrammen i sig selv kan være et resultat, hvor rammen præsenteres og illustreres via citater. Fundene er i dette tilfælde nærmere anvendt som udgangspunkt til videre undersøgelse, da resultatet af den kvalitative indholdsanalyse er undersøgt for mønstre og problematikker. Dertil fortolkes også relationerne mellem kategorierne, og ikke blot de individuelle kodninger (Schreier, 2014, s. 15).

Det endelige resultatet af den kvalitative indholdsanalyse præsenteres i Kapitel 3 og er ligeledes anvendt løbende som introduktion til og beskrivelse af problemfeltet. Der gøres fra forfatterens side opmærksom på, at det aldrig kan lade sig gøre at være udelukkende objektiv i udvælgelse af hvilke passager, der fokuseres på og behandles. Der er dog forsøgt at tage en så objektiv tilgang som muligt, hvilket også er begrundelsen for den strukturerede tilgang til kodningsprocessen og transparensen i projektet.

2.2 Teori

I dette afsnit præsenteres de designteorier, der vil udgøre den teoretiske ramme for opgaven. Her vil teorierne Design Thinking, Human Centered Design og Participatory Design blive redegjort for, da disse anvendes løsningsorienteret til det materiale, der behandles senere i opgaven. Der findes andre designteorier, der kunne have været anvendt, men netop disse er udvalgt grundet deres brugercentrerede tilgange, hvilket var en del af opgavens hypotese, at brugeren kunne være et vigtigt element i løsningen af problemstillingerne. Der vil også blive redegjort for AI, da teknologien fremgår i flere af eksemplerne i løbet af analysen. AI er derudover et bredt term, der indgår i specifikke underkategorier som eksempelvis ansigtsgenkendelse. En forudgående viden om dette er derfor nødvendig for at kunne analysere og fortolke de fremhævede problemstillinger. Teorierne anvendes til løsningsforslaget, hvor det reflekteres op imod det analyserede materiale.

2.2.1 Kunstig intelligens

Følgende afsnit består af en kort redegøre for essentielle begreber indenfor AI, da en sådan baggrundsviden er til fordel for senere analyse. Det bemærkes, at AI i denne opgave anskues ud fra en Human Centered vinkel, og skal således ikke ses som et forsøg på en teknisk redegørelse for teknologien, da dette er uden for pågældende fagområde.

I udviklingen af AI kiggede forskere på opførsel og ræsonnement hos intelligente enheder som mennesker, og udviklede algoritmer baseret på denne opførsel. Disse algoritmer er siden blevet anvendt til at løse mange problemstillinger, inklusiv medicinsk diagnosticering, fejlfinding i software, finansiel beslutningstagen, navigation i udfordrende terræn, stemme- og ansigtsgenkendelse, forståelse af tekst samt læring af individuelle præferencer. Teknologien opstod med målsætning om at skabe computersystemer i stand til at lære, reagere og tage beslutninger i komplekse, foranderlige miljøer (Neapolitan & Jiang, 2018, s. 3-4).

Indenfor AI og maskinlæring er der to grundlæggende læringstilgange: *Supervised Learning* og *Unsupervised Learning*. Den primære forskel på de to tilgange er mængden af menneskelig indblanding, hvilket uddybes i følgende afsnit.

Supervised Learning

I denne læringsproces ved AI systemet allerede, hvad det korrekte svar er, og tilpasser algoritmen, således at svaret kan blive så præcist så muligt ud fra et eksisterende datasæt. Datasættet er udarbejdet af menneskelig identificerede elementer af input og output data, hvilket algoritmen er trænet på. Et sådant AI system kan bruges til at genkende, lokalisere, kategorisere og isolere objekter, eksempelvis til at forklare priser på forskellige cykelmodeller ud fra deres karakteristikker, herunder mærke eller andre egenskaber. Systemet lærer selvstændigt ud fra prædefinerede datasæt til at genkende relevante mønstre. Da denne type kræver menneskelig intervention til at mærke dataene korrekt, har den tendens til at være mere korrekt.

Datasættene er designet til at træne eller overvåge algoritmer til at klassificere data og udfald nøjagtigt. Menneskelig involvering er dertil nødvendig for at træne algoritmens viden. Supervised Learning anvender to typer af metoder til analyse af data:

- **Klassifikation.** Problemer der anvender en algoritme til nøjagtigt at sortere testdata i specifikke kategorier. Eksempelvis når mailsystemer sorterer spam mail i en separat mappe fra indbakken. I tilfældet vil algoritmen være trænet på en base af millioner af e-mails på forskellige parametre, hvortil den vil have lært særlige kendetegn, der hvis disse er til stede, vil identificere en email som spam og således klassificere den som spam.
- **Regression.** Anvender en algoritme til at forstå relationen mellem variabler der er afhængige og uafhængige. Disse modeller er brugbare i at forudse numeriske værdier baseret på forskellige datapunkter. Denne type kan eksempelvis bruges til vejrudsigt, hvor den gennem tidligere data kan forudsige vejret i fremtiden.

Den primære forskel mellem de to typer er, at klassifikation anvendes til at klassificere værdier, eksempelvis mand eller kvinde, hvor regression bestemmer kontinuerte værdier som priser og alder (Kreutzer & Sirrenberg, 2020, s. 6-7).

Unsupervised Learning

Unsupervised Learning har til formål at finde gemte strukturer i data. Denne type har ikke foruddefinerede værdier, og må derfor selvstændigt genkende ligheder og mønstre i datasæt. Konsekvensen er, at brugeren er uvidende om disse mønstre i forvejen, og systemets viden kan derfor ligge udenfor hvad der tidligere var menneskeligt tænkeligt. Algoritmer fra

maskinlæring anvendes til at analysere og kategorisere umærkede datasæt, hvor skjulte mønstre i dataene opdages uden menneskelig involvering. Trods denne type er selvlærende, kræves der stadig menneskelig intervention for at validere udfaldene. Et eksempel på Unsupervised Learning er genkendelse af mennesker på sociale medier, som er særligt modtagelige for at tro på falske beskeder, interagere med dem positivt samt dele dem. Her er det muligt at finde ud af egenskaber for disse personer, eksempelvis at de hyppigt har kattebilleder eller er aktive på medierne i et bestemt tidsrum. Sådanne fund kan ligge udenfor hvad der før var menneskeligt muligt (Kreutzer & Sirrenberg, 2020, s. 7). Dermed består det Unsupervised i, at opdage ligheder, associationer eller lignende. En Unsupervised klyngeanalyse undersøger, hvilke elementer i en pulje af data ligner hinanden, og skaber klynger derefter, eksempelvis farver i billedgenkendelse eller lignende ord, som set med nærværende projekts ordfrekvensanalyse.

Explainable AI

I takt med at algoritmer bliver mere anvendt og nøjagtige, bliver det mere presserende at forstå, hvordan og hvorfor en forudsigelse bliver lavet. Uden forståelse er det ikke muligt at stole på forudsigelser af applikationer med AI. Termet "Explainable AI", efterfølgende XAI, har netop til formål at øge denne troværdighed, så maskinlæring fortsat kan anvendes med bedre sikkerhed (Kamath & Liu, 2021, s. 1).

Innovationen i maskinlæringsalgoritmer har resulteret i mere nøjagtige forudsigelser, men er sideløbende blevet mere kompleks: "This is an unfortunate trade-off between improved quality and transparency." (Kamath & Liu, 2021, s. 1). Det er muligt at observere og vurdere resultater for et givent sæt af inputs for en model, men det er i nogle tilfælde uden viden om det interne mellemliggende arbejde. I nogle tilfælde er det umuligt at forstå hvordan forudsigelser er lavet. Problematikken om manglende transparens i et AI-system kaldes Black Box. Fænomenet opstår, når modeller er for komplekse til at kunne fortolkes af mennesker, og er derfor svære at anvende på et ordentligt oplysningsgrundlag. Denne mangel på gennemsigtighed i forståelsen kan have alvorlige konsekvenser for troværdigheden i anvendelsen af sådanne systemer. Det kan eksempelvis være umuligt at vide, om modellens forudsigelser er forkerte, hvilket kan have fatale konsekvenser på områder som sundhed, hvor korrekt viden og faktuelle oplysninger er yderst vigtige (Kamath & Liu, 2021, s. 2).

For at imødekomme udfordringerne med Black Box fænomenet søger XAI at skabe indsigt i de beslutninger, AI skaber, herunder en forståelse af hvordan, hvornår og hvorfor en forudsigelse bliver lavet. Med XAI er det muligt at skabe større troværdighed og bedre sikkerhed i brugen af teknologi, og dermed højere mulighed for at samfund får gavn af de fordele, AI tilbyder. For at forstå XAI præsenteres fire målsætninger.

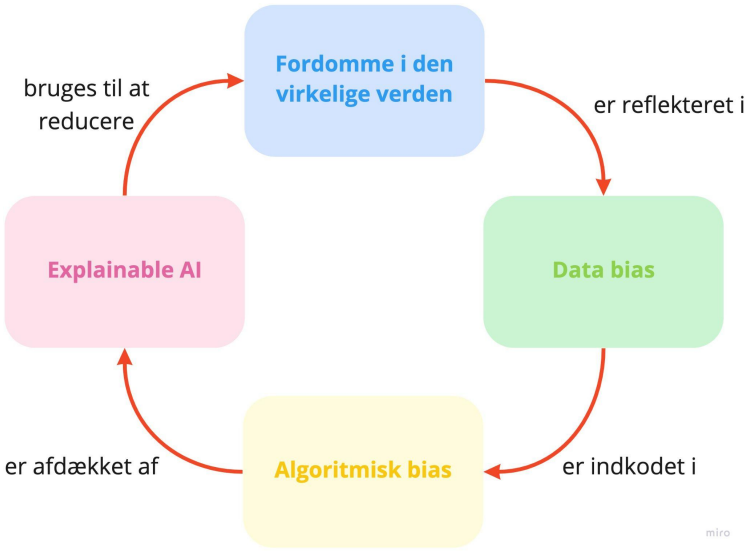
Først skal underliggende funktioner af AI være forståelig for mennesker. Den overordnede model skal være forståelig uden nødvendighed for detaljer eller forklaringer af modellens indre algoritmiske struktur. Dernæst skal en models evne til at repræsentere og formidle indlært viden være fyldestgørende på en måde, der er forståelig for mennesker. Dernæst skal det være muligt at forklare en models struktur, der er forståelig for mennesker gennem simple termer. Da fortolkning er subjektiv, skal der her tages højde for publikum og kontekst, da forståelsen skal passe hertil. Det kunne eksempelvis være ved at bruge forskellige forklaringer til tekniske eksperter og brugere, der ikke har teknisk viden. Sidst skal en models indre struktur og algoritme være transparent, således at forudsigelserne er forståelige. Det hjælper til at adressere hvorfor en model fungerer, som den gør. Dette er eksempelvis vigtigt i sundhedstilfælde, hvis en model beslutter hvilke patienter skal have hjælp først (Kamath & Liu, 2021, s. 3-4). Målsætningerne har det tilfælles, at de stræber efter mere troværdig AI gennem oplysning og gennemsigtighed.

Med AI følger praktiske, etiske, filosofiske og retfærdigheds problemstillinger, som XAI kan være med til at adressere og i bedste udfald løse. Kamath & Lui (2021) præsenterer følgende ni formål med XAI:

1. **Oplysningsevne**. AI gør det i praksis nemmere at tage beslutninger. AI-modeller er designet til at opnå specifikke kvantitative mål, men i realiteten matcher disse mål sig ikke altid med den oprindelige hensigt. Konsekvenserne for denne uoverensstemmelse kan være katastrofale, og XAI er derfor nødvendig til at kunne muliggøre en evaluering af, om eller hvornår en forudsigelse kan være forkert eller kontraproduktive i forhold til formålet (Kamath & Liu, 2021, s. 7).
2. **Troværdighed.** En AI er kun så troværdig, som den kan forklares. Det er nemmere at stole på forståelige algoritmer, som ikke er tilbøjelige til at misrepræsentere. Troværdighed er svært at måle, og der er forskel på at stole på en AI model og på dens troværdigheden. Hvis en person stoler på en AI model, er denne troværdighed drevet

af holdninger, hvor troværdigheden i selve modellen måles i forhold til om den opnår dens formål (Kamath & Liu, 2021, s. 8-9). Hvis en AI model forudser, at en gruppe patienter har førsteret til behandling, kan troværdigheden være høj i form af, at modellen har opnået dens formål om at forudsige patientrisiko. Det betyder ikke nødvendigvis, at vi som patienter eller lægfolk stoler på, at resultaterne er korrekte.

3. **Retfærdighed.** Det samfundsmæssige ansvar om retfærdighed er en vigtig målsætning for AI, hvor XAI muliggøre at identificere eksisterende bias i en model. Retfærdighed beskrives her som en upartisk og ligeværdig behandling samt fravær af favorisme og diskrimination. Følgende model er en oversættelse af Kamath & Lius (2021) model, der illustrere forholdet mellem bias og XAI.



Billede 2. Bias og XAI

Modellen illustrerer, hvordan fordomme i den virkelig verden reflekteres ind i data, der deraf bliver biased, hvilket dernæst indkodes i algoritmer. Her kan XAI anvendes til at afdække og reducere de opståede bias. Den cirkulære proces viser hvordan og hvornår i processen, XAI anvendes til at bekæmpe bias og dermed afhjælpe de problematikker, som bias medfører.

4. **Transparens.** XAI er kendetegnet ved transparens, der understøtter forståelsen af hvordan beslutninger tages. Det er derfor en essentiel del af at øge troværdighed. Transparens betyder dog ikke nødvendigvis retfærdighed eller skaber de rette forklaringer (Kamath & Liu, 2021, s. 9). Hvis en AI vurderer recidiv i kriminalsager på baggrund af historisk data der er gennemsigtigt, betyder dette ikke nødvendigvis, at

vurderingen er korrekt. Hvis den historiske data er bias i sig selv, og dette ikke bliver vurderet, vil historien gentage sig.

5. **Kausalitet.** En grundlæggende begrænsning ved AI er manglen på årsagssammenhænge. Teknikker for modellerne anvender korrelation, men ignorerer faktorer som tid. XAI har i stigende grad til formål at identificere årsagssammenhænge i dataene, som kan bruges til at forklare og udforske årsag og effekt. Hvad er årsagen til beslutningerne, og hvilken effekt har de.
6. **Overførbarhed.** En model trænet på én opgave kan generaliseres og bruges som udgangspunkt for en anden, og det kan være en fordel at anvende eksisterende viden til at bygge ovenpå. XAI kan anvendes til at vurdere hvornår og hvorledes en model er overførbar, da den søger at identificere og forstå hvilke begrænsninger der eksisterer i en model, der kan afgøre overførbarheden.
7. **Pålidelighed.** Pålidelighed afhænger af graden af en models troværdighed. Pålidelighed og stabilitet er karakteristikker, XAI søger at opnå, således at en AI model tager sammenlignelige beslutninger i sammenlignelige tilfælde.
8. **Tilgængelighed.** Tilgængelighed af AI applikationer for ikke-tekniske personer er et vigtigt element i at øge popularitet og adoption af teknologier. XAI kan gøre det nemmere for ikke-tekniske mennesker at anvende og forstå AI gennem facilitering af viden og forståelse af kompleksitet.
9. **Privatliv.** Sikkerhed og privatliv er i øget fokus i AI applikationer, hvilket XAI spiller en rolle i at sikre. Med XAI er det nemmere at evaluere hvorvidt rettigheder for privatliv bliver overholdt gennem krypteret repræsentationer eller algoritmer, hvor læsbare formater konverteres til kodet formater (Kamath & Liu, 2021, s. 10). Det kan eksempelvis være bankoplysninger eller passwords der sikres gennem kryptering.

AI kan have samfundsmæssige fordele og være med til at bygge en bedre verden, hvor XAI en vigtig del i at facilitere en bedre adoption af AI applikationer ved at understøtte og adressere problemstillinger som retfærdighed, bias, verificering, sikkerhed og ansvarlighed (Kamath & Liu, 2021, s. 10-11). XAI tilbyder forskellige forklaringsvarianter, der hjælper med at forstå komplekse systemer. Følgende fem typer er forklaringer, XAI tilbyder for bedre forståelse:

1. **Globale forklaringer.** Søger at svare på hvordan en model fungerer ved at forklare, hvordan en model har fundet frem til dens forudsigelser. Svaret kan komme i form af

visuelle tabeller, matematiske formularer eller grafer med modeller. Disse forklarer algoritmen som en helhed, og inkluderer eksempelvis svar på hvilken træningsdata der er anvendt og advarsler om svagheder i algoritmen.

2. **Lokale forklaringer.** Søger at svare på hvorfor en model er kommet frem til dens forudsigelser ved at forklare en enkelt forudsigelse ad gangen, der kan begrundes med en specifik egenskab af dataene eller algoritmemodellen.
3. **Kontrastive forklaringer.** Hjælper med at forstå hvorfor en model laver én forudsigelse kontra en anden. Anvendes ofte sammen med spørgsmål om hvorfor, for at forstå modellens forudsigelser og forventede resultater (Kamath & Liu, 2021, s. 12).
4. **Hvad-nu-hvis forklaringer.** Hvad-nu-hvis spørgsmål er brugbare til at understøtte forståelse af relationen mellem forudsigelser og egenskaber. Anvendes ved forandringer i modellens input og parametre. Eksempelvis kan det vurderes, hvilke udfald det ville have, hvis datasættet for en model blev ændret.
5. **Kontrafaktiske forklaringer.** Fortæller hypotetiske ændringer i input eller parametre i en model, og søger at svare på hvordan et ønsket udfald kan opnås. Dette gøres ved at beskrive de mindste ændringer til en model, uden at behøve at forstå den indre struktur.
6. **Eksempelbaserede forklaringer.** Gennem konkrete eksempler kan en models opførsel og underliggende datafordeling forklares. Ofte præsenteres lignende tilfælde fra en model der vil have lignende udfald (Kamath & Liu, 2021, s. 12-13).

De belyste forklaringsmuligheder illustrere hvordan XAI kan anvendes til at forstå kompleksiteten ved AI, til således at skabe troværdighed og tryghed i brugen af teknologien.

For at kunne evaluere kvaliteten af XAI, er det nødvendigt at definere forklaringernes egenskaber og dernæst kortlægge forskellige teknikker til disse. Teknikkerne skal give klarhed i form af entydighed og enkelthed i form af simple forklaringer. Derudover skal de tilbyde bredde i form af, at samme forklaring skal kunne anvendes over en række observationer, fuldstændighed gennem tilstrækkelig information og sidst forsvarlighed som indikator for, hvor troværdig og korrekt en forklaring er (Kamath & Liu, 2021, s. 304). En forklaring kan klassificeres i følgende kategorier:

1. **Rationale.** Hvorfor-delen af maskinlærings beslutningstagen. Forklaringen tillader brugeren at forklare hvorfor en beslutning er fejlagtig eller give begrundelse der understøtter beslutningen.
2. **Ansvarlighed.** Hvem der er involveret i beslutningerne af forskellige trin i en maskinlæringsproces, eksempelvis fra data til model. Denne type af forklaring bringer ansvarlighed og sporbarhed.
3. **Data.** Data har en central rolle i beslutningstagning. Dataforklaringer fokuserer på hvilke data der er brugt til træning, validering og test af modeller.
4. **Retfærdighed.** Fokuserer på at sikre, at der ikke er bias eller diskrimination i maskinlærings beslutningsprocesser. Et kritisk område i at øge troværdigheden for AI og maskinlæringssystemer, og et fokus for denne opgaves undersøgelse.
5. **Sikkerhed** og præstation. Fokuserer på at maksimere robusthed, pålidelighed og præcision i de beslutningstagende trin.
6. **Indvirkning.** Understreger den indvirkning, maskinlæringssystemer og beslutninger har på forskellige niveauer, fra individer til samfund (Kamath & Liu, 2021, s. 304).

Endvidere fordeles metoder til evaluering af XAIs kvalitet i tre kategorier:

- **Applikationsbaseret evaluering.** I denne tilgang vurderer fageksperter eller slutbrugerne af systemet, i hvor høj grad forklaringen understøtter dem i beslutningstagen.
- **Menneskebaseret evaluering.** I modsætning til forrige evaluering vurderer her en stor gruppe ikke-eksperter forklaringskvaliteten i en menneskebaseret evaluering. Tilgangen er billigere at udføre, men kan have betydelige variationer afhængig af opgavens størrelse og type.
- **Funktionalitetsbaseret evaluering.** I denne type evaluering fungerer en kvantificerbar metrik som fuldmagt for forklaringsevalueringen. Fuldmagten er baseret på en formel definition af mulighed for fortolkning (Kamath & Liu, 2021, s. 305).

Metoderne er brugbare i relation til at bedømme hvorvidt forklaringen af et system er tilstrækkelig til at kunne vurdere systemets troværdighed og brugbarhed. Da beslutninger af AI ofte kan være uforståelig, er det presserende at fænomener som XAI bliver anvendt til at skabe bedre forståelse for, hvordan og hvorfor disse teknologier bliver anvendt. Dette vil illustreres løbende i analyseafsnittet.

2.2.2 Human-Centered Design

Human-Centered Design, efterfølgende HCD, anvendes blandt andet i AI, og bliver beskrevet som en tilgang til design og udvikling af systemer, der har til formål at lave interaktive systemer mere brugervenlige ved at fokusere på systemets brug og tilføje menneskelige faktorer samt viden og teknikker om anvendelighed. HCD kan beskrives med følgende principper:

- Inddragelse af tværfaglige færdigheder og perspektiver.
- Eksplicit forståelse af brugere, opgaver og miljøer.
- Brugercentreret evaluering.
- Hensyn til hele brugeroplevelsen.
- Involvering af brugere gennem design og udvikling.
- Iterativ proces (Giacomin, 2014, s. 608).

Endvidere implicerer synet på HCD at enhver design aktivitet er identifikationen af den mening, et givent produkt, system eller service bør tilbyde brugere. HCD er i dag baseret på brugen af teknikker der kommunikerer, interagerer, som har empati og stimulerer brugere ved at opnå en forståelse af behov, ønsker og oplevelser, der ofte overgår brugernes egne indsigter. Med andre ord kender brugerne ikke altid selv deres behov. Ved praktisering af HCD opnås resultater der på alle måder er intuitive, forstået på den måde at produktet, systemet eller servicen kan manipuleres nemt og med det samme. (Giacomin, 2014, s. 609-10).

En Human-Centered designer er i dag en relativt transparent figur, der ikke forsøger at pålægge præferencer på et projekt, men i stedet simulerer, formidler og oversætter viljen af de involverede. Dette indikerer en høj fleksibilitet i HCD. HCD værktøjer kan klassificeres baseret på intentionen i anvendelsen, hvilket inkluderer fakta om mennesket, som antropologisk og kognitiv data. Sådanne informationer tilbyder faktuelle udsagn om muligheder og begrænsninger af mennesker, og definerer således grænser for HCD processen. Det kan eksempelvis være vejledende i hvordan et system skal fungere baseret på de pågældende brugeres muligheder og begrænsninger.

HCD værktøjer kan være sprogligt baserede, såsom etnografiske interviews eller spørgeskemaer. Som tidligere belyst er det ikke altid muligt at indhente informationer gennem bevidsthed, og derfor er teknikker som deltagerobservation i stigende brug. Det kan

også være simulering af intuition gennem co-design og prototyper, hvilket fordyber brugerne i mulige løsninger ved at give dem mulighed for at eksperimentere og interagere med et produkt og udforme personlige perspektiver og meninger (Giacomin, 2014, s. 614-615).

Med HCD illustreres overordnede principper og tilgange til design af teknologi, hvilket udgør et udgangspunkt for teorierne Design Thinking og Participatory Design, der kan siges at udspringe af HCD. Det vil løbende blive illustreret, hvordan de tre teorier relaterer til hinanden. Fælles for dem er den brugercentrerede tilgang, hvor HCD kan siges at være den mest overordnede, og derfor udgangspunktet for afsnittet.

2.2.3 Design Thinking

Følgende afsnit består af en kort redegørelse for Design Thinking, efterfølgende DT, med udgangspunkt i Razzouk & Shutes artikel (2012) der beskriver processen. Sidst anvendes en model fra Interaction Design Foundation, der tilbyder en simplificeret udgave af DT til anvendelse i analyseafsnittet (Razzouk & Shute, 2012).

For at være succesfuld i nutidens højteknologiske og globalt konkurrerende verden er det nødvendigt at udvikle og anvende færdigheder, der ikke førhen har været nødvendige. En af disse færdigheder er DT, som ofte anskues som en analytisk og kreativ proces der involverer muligheder for at eksperimentere, indsamle feedback og redesigne. Det involverer kreativ tænkning i skabelsen af løsninger på problemstillinger, hvor tilgangen har modtaget en del opmærksomhed i ingeniør-, arkitektur- og designfag grundet muligheden for at ændre menneskelig problemløsning (Razzouk & Shute, 2012, s. 330-331).

DT refererer til hvordan designere ser og tænker. Det er en iterativ proces hvor designere kigger på koncepter eller idéer til problemløsninger, drager forbindelser mellem idéer og løsninger og ser på det, der er blevet tegnet, prototyper eller lignende, til at informere videre design. Design kan anskues som en situeret handling hvor designere opfinder designproblemer på en måde, der er situeret i det miljø, der designes i. Her findes en tovejs korrelation mellem uventede opdagelser og de opfundne problemstillinger. Uventede opdagelser er tilfælde, hvor en designer opdager noget nyt i et tidligere tegnet element i et løsningskoncept, hvilket kan blive drivkraft i opfindelsen af nye problematikker eller krav. Samtidig kan opfindelser have tendens til at forårsage nye, uventede opdagelser. Dette

understreger vigtigheden af at skifte mellem forskellige aktiviteter i en designproces, da der undervejs kan dukke nye fund op, som kan forbedre projektet. Eksempelvis kan en sketch føre til nye opdagelser, som hænger sammen med opfattelsen af det pågældende designproblem. Ordet 'Thinking' relaterer til forskellige former for tænkning, der kan observeres i en designproces. Design starter ud med en uklar idé om et produkts udseende og funktionalitet, hvilket udspringer af designerens viden. Med tiden transformeres denne idé til en komplet produktidé. Tænkning involverer også sketching og prototyper som er metoder til at konkretisere idéer. Disse klargøre karakteristikkerne af et produkt, hvilket kan udgøre grundlag for processen. Sidst kan det at sætte ord på idéer endvidere hjælper med at klargøre og uddybe disse (Razzouk & Shute, 2012, s. 334-335).

Designprocessen er karakteriseret ved at være en iterativ, undersøgende og somme tider kaotisk proces, der starter abstrakt og efterhånden definerer produktspecifikationer. Den følger en cyklus af gensidige justeringer mellem specifikationer og løsninger, indtil en endelig løsning er fundet. Tre nødvendige processer er blevet beskrevet i DT:

1. **Forberedelse.** I denne proces skal designerne beslutte et fokus for, hvad der er relevant. Her udvikles specifikationer og begrænsninger, genfortolkninger af idéer, visualisering, problemformulering og lignende.
2. **Assimilation.** Her skabes mening med den foreslåede løsning, data og observationer gennem feedback fra eksperimentering med prototyper.
3. **Strategisk kontrol.** Her tages beslutninger om hvilken idé, der skal undersøges nærmere, eller hvilke begrænsninger der skal revurderes (Razzouk & Shute, 2012, s. 336-337).

Lignende modeller af DT er opstået siden teoriens oprindelse baseret på forskellige måder at anskue designsituationer på, med anvendelse af modeller fra design metodologi, psykologi, uddannelse m.fl. Tilsammen udgør de en strøm af undersøgelser, der skaber en varieret forståelse af menneskets komplekse realitet. DT anvendes ofte til at identificere nye måder at behandle problemer på i mange professioner, særligt informationsteknologi (Dorst, 2011, s. 521). Til denne opgave præsenteres en model fra Interaction Design Foundation, der tilbyder en praktisk, metodisk model, der repræsenterer egenskaberne ved DT.

Modellen består af fem stadier, der ikke nødvendigvis følges som en lineær proces. Processen er fleksibel i praksis, hvor eksempelvis testfasen kan lede til nye indsigter, der leder til en ny

idegenerering i form af brainstorming eller udvikling af en ny prototype. Ved at bruge DT arbejdes der iterativt, og det er derfor en oplagt proces at anvende til komplekse problemstillinger, hvor der er flere involverede parter. Det understøtter HCDs princip om den iterative arbejdsproces. Fordelen ved en sådan tilgang er, at information fra ét stadie kan videre- eller tilbageføres til et andet stadie, og information bliver dermed løbende anvendt til forståelse af problem og løsning samt til redefinering af problemstillingen. Processen systematiserer samtidig de fem stadier, der kan forventes i et designprojekt, hvor alle projekter med denne tilgang vil involvere en udgave af stadierne relevant for den specifikke problemstilling (Interaction Design Foundation). Således kan DT anvendes som et rammeværktøj til udvikling af teknologiske løsninger, hvor den vil vejlede udviklingen på en måde, der hele tiden har brugeren i fokus.

De fem stadier består af følgende:

1. Emphasise handler om at tilgå en empatisk forståelse af det givne problem, ved at forstå konteksten. Dette kan gøres ved at konsultere eksperter, for at finde ud af mere om et område gennem observation. Det sker også gennem engagement med brugere og andre involverede, for at forstå deres oplevelser og motivation. Derudover er det vigtigt at involvere sig i relevante fysiske miljøer for at skabe bedre indsigt i de involverede problemstillinger. Empati er uundgåelig for at kunne gå brugercentreret til et designproblem, da det tillader designere at tilsidesætte egne fordomme om verden og få indsigt i brugerne og deres behov. Informationer der indhentes i dette stadie bruges endvidere til at kunne definere en problemstilling, og skaber en basis for at opnå forståelse for brugere og deres behov. Stadiet hænger unægteligt sammen med HCDs principper om at opnå brugerforståelse gennem empati samt indsigt i miljø og opgaver.

2. Define handler om at sammensætte de informationer, der tidligere er indsamlet til at definere en problemstilling ved hjælp af de udførte observationer. Der opfordres til at involvere brugerne i denne definering, hvilket er et HCD princip, da det sætter brugeren i centrum. Det vil samtidig være med til at sikre, at designeren ikke pålægger projektet egne præferencer samt fjerne bias. Stadiet understøtter efterfølgende idegenerering ved at etablere funktioner, egenskaber og andre relevante elementer, der kan være en del af løsningen.

3. Ideate handler om at idegenerere, herunder afprøve forskellige løsninger og opnå ny viden. Under dette stadie anvendes den brugerforståelse, der er udviklet i de forrige stadier til at

danne løsningsidéer. Metoder som brainstorming anvendes til at stimulere fri tænkning og udvide problemområdet. Målet er at danne så mange idéer og løsninger som muligt samt afprøve disse, for at finde den bedste måde at løse problemet på.

4. Prototype. En praktisk tilgang anvendes til udvikling af prototyper. Disse kan være mange, billige, nedskalerede versioner af produktet eller specifikke egenskaber ved produktet, der kan bruges til at undersøge de problemløsninger, der blev fundet ved forrige stadie. Prototyperne kan testes indenfor designteamet eller med eksterne slutbrugere. Sidstnævnte er at foretrække efter HCDs princip om brugerinvolvering. Stadiet er eksperimenterende, hvor målet er at identificere de bedst mulige løsninger for hvert problemområde. De nye løsninger kan dernæst implementeres i nye prototyper, hvor disse kan testet og evalueres ud fra brugernes oplevelser. Efter dette stadie vil designteamet have en bedre forståelse for den givne problemstilling samt hvordan rigtige brugere vil opleve produktet. Ved at inddrage brugerne i stadiet understøttes HCDs formål om at skabe intuitive produkter, da det gennem brugerne opdages, hvad der fungerer.

5. Test. Sidste stadie er testfasen, hvor produktet bliver testet med de bedste løsningsforslag identificeret i forrige fase. Da DT er iterativ, vil resultaterne fra dette stadie ofte anvendes til at redefinere problemstillingen og give information om brugeren, brugbarheden af produktet og lignende, der kan anvendes til redesign af løsning. Test kan desuden foregå i alle stadier af processen, da flere test vil lede til flere opdagelser og bedre løsninger. Test anbefales derfor at blive udført flere gange undervejs, hvilket yderligere understøtter HCDs princip om iteration (Interaction Design Foundation).

2.2.4 Participatory Design

Participatory Design, efterfølgende PD, har tradition for at forpligte sig til at sikre, at brugerne af et produkt spiller en central rolle i designet. Det betyder at mennesker, hvis liv vil blive påvirket af et produkt, er co-designere af selv samme. Det defineres endvidere som en proces, hvor gensidig læring mellem flere deltagere understøttes i et kollektivt samarbejde. Designernes mål er at lære brugernes realiteter, og brugernes mål er at lære at udtrykke deres behov og samtidig forstå relevante teknologiske midler, der kan hjælpe med at opnå disse behov (Robertson & Simonsen, 2012, s. 2, A). Det første element i teorien, 'Participatory', beskrives som en ideologi hvor stakeholders, særligt brugere, udviklere og planlæggere i samarbejde skaber eller justerer teknologier på en måde, der passer bedre til brugernes behov.

Derudover er det en ideologi, der henviser til spørgsmål om politik, etik, demokrati og indflydelse (Bannon & Ehn, 2012, s. 41). Brugerinvolvering opfordres på samme måde i DT, hvor PD giver et bedre indblik i fordelene ved sådant engagement.

Fundamentet i PD bygger på at give en stemme til dem, der skal bruge teknologien, uden at de behøver at kunne det tekniske sprog. Dette opnås gennem brugerinteraktioner med prototyper og andre værktøjer der repræsenterer det udviklende system og fremtidige praksisser. Derudover giver en gensidig læring mellem bruger og designer brugeren mulighed for at italesætte behov, da de gennem læring opnår en forståelse for, hvad der er muligt. Med 'designere' menes der de, der professionelt er ansvarlige for den givne udvikling af et projekt, hvor 'brugere' er de deltagere, der i sidste ende vil interagere med det færdige produkt (Robertson & Simonsen, 2012, s. 2-3, A). Dette understøttes i DTs fjerde fase, om test af prototyper med slutbrugerne. Derudover også HCDs princip om at opnå intuitive løsninger, da det kun er muligt gennem inddragelse af slutbrugere at teste, om en løsning er intuitiv at anvende.

Designprocessen har en etisk underliggende funktion, da den anerkender et ansvar for design til den verden, den skaber og livet for dem, den påvirker. Det er et samarbejde mellem dem, der designer, og dem, der skal bruge designet, og tager dermed ansvar for at sikre relevante, oplagte løsninger (Robertson & Simonsen, 2012, s. 5, A). Ansvarlighed er også en del af XAI, der søger gennemsigtighed i, hvem der er involveret i beslutninger, for dermed at bringe ansvarlighed.

'Participant' i PD dækker over involvering af brugere i designløsningen, der rækker sig udover blot at interviewe disse. Det betyder anerkendelse af brugere som co-designere, ved at give medbestemmelse og del af designprocessen. Processen involverer brugerens perspektiv på problem og løsning, og den form for involvering kræver troværdig og fortrolig relation mellem alle involverede. I tilfælde med AI kræver en sådan proces tilgængelighed gennem XAI, der sikrer forståelse for ikke-tekniske personer. Det kræver desuden en overvejelse af magtfordeling i forhold til medbestemmelse af designbeslutninger samt hvordan designprocessen skal se ud for at kunne have flere deltagere (Robertson & Simonsen, 2012, s. 5, A).

PD har bevist forøget kvalitet i produkter samt øget kvalitet i processen, hvor en involvering af hverdagsviden for et givent område skaber bedre kvalitet, service og funktionalitet. Ved brug af eksempelvis prototyper gøres det muligt for "almindelige arbejdere" at bruge deres praksiserfaring i deltagelsen i designprocessen, og det øger deltagernes engagement og fordrer bedre kommunikation og fælles forståelse mellem de forskellige stakeholders (Robertson & Simonsen, 2012, s. 6, A).

Det ovenstående kapitel om, HCD, DT og PD samt dele af XAI, viser hvordan de fire teorier kan anvendes i fællesskab om samme intentioner. Fælles er brugeren som central rolle i designudvikling, der ifølge teorierne vil føre til bedre løsninger.

Kapitel 3 - Analyse og resultater

Dette kapitel indeholder opgavens analyser og resultater. Analyserne består af en gennemgående beskrivelse af tendenser opdaget i litteraturen, og dykker således ikke nødvendigvis ned i hver enkelt artikel. Formålet med analysen er at afdække problemområdet for at kunne besvare problemformuleringens spørgsmål, og med en sammenfatning af forståelse af problemfeltet samt den tidligere belyste teori at kunne komme med et løsningsforslag.

3.1 Analyse af problemidentificerende materiale

Dette afsnit indeholder den beskrivende del af opgavens analyse, som er metodisk baseret på det tidligere metodeafsnit (2.1.4) om kodningsrammen, hvor der her, som beskrevet, er anvendt en modificeret udgave. Til dette anvendes den problemidentificerende del af det udvalgte litteratur til indledningsvis at redegøre for problemfeltet. Der præsenteres to kategorier udledt af kodningen, hvor den første skaber overblik over etiske udfordringer i design og udvikling af teknologi, og sidste kategori fokuserer på det udvalgte bias.

3.1.1 Etiske udfordringer i design og udvikling af teknologi

Følgende afsnit er udledt af problemformuleringens første spørgsmål, der er en redegørelse for, hvilke bias relaterede problematikker der eksisterer indenfor design og udvikling af teknologi. Kategorien fungerer således som problemdefinerende for, hvilke bias problematikker der er belyst i den problemidentificerende litteratur. Kategorien indeholder 61 koder fordelt på 11 artikler. Som tidligere belyst blev der ved en ordfrekvensanalyse af de 11 artikler opdaget, at begreber i området AI var mest forekommende. Dette vil også komme til udtryk i denne analyse, hvor teknologien er en stor del af de beskrevne problemstillinger.

Som det vil blive illustreret i gennemgangen af det problemidentificerende materiale er der flere til dato stadig uløste etiske udfordringer med udvikling og brug af teknologi. Teknologi bliver anvendt over hele verden i mange kritiske områder, der dagligt påvirker mennesker

både direkte og indirekte. Derfor er det vigtigt, at der bliver sat fokus på disse udfordringer, så de kan blive objektiv for undersøgelser og løsninger.

Beslutninger med kunstig intelligens

AI opererer ved, at teknologien tager beslutninger på baggrund af data, der er tilføjet til denne. Det er bredt anvendt af virksomheder, regeringer og andre organisationer, der anvender teknologien til at tage beslutninger, der har vidtrækkende indflydelse på individer og samfund. Disse beslutningers formål er at tilbyde indsigt eller løsninger til problemer, og kan have indflydelse på alle mennesker hvor som helst, og når som helst. Med det følger også en risiko for konsekvenser som at blive nægtet et job eller medicinsk behandling (Ntoutsi et al., 2020, s. 2). Leavy, Siapera & O'Sullivan (2021) beskriver i denne sammenhænge, hvordan bias indlejret i den anvendte data har potentiale til at resultere i en videreførelse af social ulighed. Dette er illustreret tidligere i billede 2 (afsnit om kunstig intelligens, modellen), som viser hvordan bias data indlejres i algoritmer, og deraf vil fortsætte disse bias. Denne bekymring er en relevant problematik der kan have store konsekvenser for i forvejen marginaliserede målgrupper, og diskriminerende indflydelse af AI-baserede beslutninger til særlige befolkningsgrupper er observeret i en lang række tilfælde (Ntoutsi et al., 2020, s. 2). Problemet kan derfor siges at være dobbelt problematisk, fordi det viderefører og forstærker uligheder, der eksisterer i forvejen, og derfor rammer særligt sårbare målgrupper. Det kan være som følge af historisk ulighed, der sker på baggrund af historisk bias i repræsentation, klassifikation og datarelateret kategorisering af mennesker (Leavy et al., 2021, s. 695). Hvis der eksempelvis historisk set er forsket mere i etnisk hvide menneskers sundhed, vil der opstå misrepræsentation af andre etnicitetsgrupper, som deraf vil opleve ulige behandling i sundhedsrelationer. Dette nødvendiggøre en presserende reform af den praksis, der omhandler udviklingen og brugen af AI, herunder etiske og regulatoriske rammer, da konsekvenserne ellers kan være: "[...] that decades of advances in human rights and civil liberties may be undermined." (Leavy et al., 2021, s. 695).

Menneskerettigheder og diskrimination

Menneskerettigheds- og borgerrettighedsbevægelser handler om rettighed til selvbestemmelse og frihed fra at blive klassificeret, fremstillet stereotypt og forskelsbehandlet. Udfordringerne med systemer der anvender AI er netop evnen til at fastholde diskrimination, der modstrider disse essentielle rettigheder. Denne modstridenhed har vakt årtiers debat om social retfærdighed og overvågning til live med den stadigt stigende

brug af AI. Tilfælde af etnisk- og kønsdiskrimination afsløres i brugen af AI, der afspejler og i nogle tilfælde endda forværrer diskriminerende adfærd, som borgerrettighedsbevægelser har kæmpet imod. Når AI bruges til at klassificere individer i grupper, generalisere på baggrund af antagelser af disse med forskelsbehandling som resultat, fremhæver behovet for grundlæggende etiske retningslinjer (Leavy et al., 2021, s. 696). AI er afhængig af data, der genereres af mennesker eller indsamlet via systemer lavet af mennesker. Det kan derfor aldrig siges, at være fuldstændig objektiv, da de bias, der uundgåeligt eksisterer i alle mennesker, vil blive videreført og eksistere i den data, der bliver anvendt til teknologien. Algoritmer vil dermed reproducere og i nogle tilfælde forstærke eksisterende uligheder og diskriminationer, hvor særlige grupper vil blive forfordelt. Dette resulterer som regel i såkaldt institutionel bias, hvor praksis og procedurer af institutioner, som banker og sundhedsvæsen, opererer på en måde, hvor særlige målgrupper bliver forfordelt (Ntoutsi et al., 2020, s. 3).

Datadreven objektivitet

Maskinlæring som en samling af objektiv viden er et debatteret område. Hvor store datasæt kan love datadreven, objektive sandheder, er den faktisk sandhed i mange tilfælde en anden. Beviser på det modsatte kan findes i menneskelige beslutninger omkring hvordan befolkningen repræsenteres i data, samt hvordan disse datasæt er indsamlet. Kontekst er med andre ord en vigtig faktor, som maskinlæring ikke nødvendigvis inddrager eller gør opmærksom på. Leavy et al. (2021) belyser i deres artikel et forsøg med softwaren COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), der forudser recidiv i forbindelse med kriminalitet. Her forsøges at imødekomme og fjerne bias, hvor resultatet implicerer, at det ofte er umuligt at eliminere bias fuldstændigt fra AI systemer. Det skyldes blandt andet, at værktøjets forudsigelser er baseret på en model for politiarbejde der er kritiseret for at være institutionelt racistisk, hvilket umuliggører separationen mellem AI systemer og den bredere, sociale kontekst (Leavy et al., 2021, s. 697).

Et andet eksempel er Googles reklameværktøj, der viste at målrette færre reklamer til højt betalte jobs til kvinder, end der blev målrettet til mænd (Ntoutsi et al., 2020, s. 2). Det er derfor vigtigt at anerkende data som noget, der hverken er neutralt eller objektivt, men som et værktøj hvis virkning og udfald er betinget af, hvordan det anvendes og udvikles. At forstå de beslutninger der er involveret i databehandling som politisk påvirket kræver en ny, regulatorisk proces og en kritisk refleksion om hvorvidt AI kan anvendes neutralt (Leavy et al., 2021, s. 695).

Repræsentation

En af de belyste årsager til de belyste udfordringer med AI handler om mangelfuld repræsentation på området: "[...] the dominant group whose values are likely to be captures are likely to be straight, white, cisgender, college-educated men." (Leavy et al., 2021, s. 699). Når en dominant gruppe står for at designe strukturen for datasæt, skabe dataene og designe de tilhørende etiketter vil denne dominante gruppes bias' uundgåeligt indlejre sig i disse data og videreføres gennem maskinlæringen, hvor disse data anvendes. Et eksempel på en konsekvens af denne form for indkodet social magtstruktur er en kritik af Twitters brug af algoritme til billedbeskæring, der hovedsageligt beskar kvinders samt mennesker med mørkere hud ud af billeder. Problemet opstod, fordi algoritmen fungerer med et interessefokus, der ekskluderede særlige målgrupper (Leavy et al., 2021, s. 699). Data kan siges at være et direkte resultat af et særligt interessefelt, der har forskellige påvirkninger, heriblandt køn og etnicitet.

Repræsentation har dermed direkte eller indirekte indflydelse på udfaldet af teknologi. Det er ikke kun vigtig i det data, der bruges i teknologier, men også i designet af teknologien. Overrepræsentation af mænd i design af teknologi kan på en diskret måde underminere årtiers fremgang i ligestilling for kønnene. AI lærer af den data, den bliver præsenteret for, og hvis denne data er ladet med stereotypiske koncepter af eksempelvis køn, kan dette resultere i at teknologien foreviger disse bias. Ydermere er ledende forskere indenfor det fremvoksende felt, der adresserer bias i AI, primært kvinder, hvilket antyder, at dem, der er mest påvirket af bias har større tendens til at se, forstå og forsøge at behandle problematikken. Derfor er repræsentation i udviklingen særlig vigtig hvis problematikkerne skal løses (Leavy, 2018, s. 14).

Stereotyper og terminologi

Kønsbias er en fremhævet problemstilling, og kan forstås i de termer, der anvendes til beskrivelse af forskellige køn. Et eksempel er ordet "familiefar", hvor der ikke tilsvarende bruges "familiemor", eller "karrierekvinde" hvor der ikke bruges "karrieremand". Denne måde at bruge termer på illustrerer en identificering af eksistens der modsiger sociale forventninger. Derudover bliver term som "han", "ham", og "mand" ofte anvendt til at beskrive alle køn. I relation til maskinlæring vil disse kønsbias unægteligt være en del af teknologiens træningssæt, og resultater vil dermed reflektere forældede måder at referere til køn på og vedligeholde stereotyper (Leavy, 2018, s. 15).

Eksempler på stereotypforstærkende teknologi er stemmeassistenter, der ofte anvender kvindelige stemmer som standard i underdanige roller, og samtidig ekskluderer marginaliserede grupper som ikke-binær kønsidentitet som en valgmulighed. Undersøgelser med robotter fandt, at brugere tilskrev det kvindelige eller mandlige køn til robotter baseret på deres funktion. Her blev robotter tildelt sikkerhedsarbejde set som mandlige, hvor den samme robot der var sat til at give vejledning til eksempelvis retning blev anset som kvindelig (Danielescu, 2020, s. 23). Selvom en designer ikke eksplicit tildeler personlighed til teknologi, vil brugere tildele en til det på baggrund af sociale normer og forventninger, da vi har tilbøjelighed til at menneskeliggøre ikke-menneskelige objekter. Danielescu (2020) beskriver, at frem for at videreføre skadende stereotyper har disse teknologier oplagt mulighed for at udfordre dem og promovere inklusion og diversitet. Hertil nævnes, at undersøgelser har vist, at brugere ikke kan lide stemmeassistenter uden klare kønsmæssige markører eller hvis der opstår uoverensstemmelser mellem stemme og personlighed. Det skaber en udfordring, hvis teknologien skal være med til at fremme den inkluderende dagsorden, og kræver mere undersøgelse hvis det skal lykkes. Fremgang i ligestilling afspejler sig ikke i nyere teknologi, hvilket åbner op for at kunne udfordre stereotyper og måske løsne blandt andet socialt konstruerede kønsroller og ændre brugerens opfattelse over tid (Danielescu, 2020, s. 22).

Finansielle institutioner

AI bliver også i et vidt omfang anvendt i finansielle institutioner, hvor det tilbyder fleksibilitet, lavere omkostning, bekvemmelighed og personaliseret service. Kigges der på teknologien i dette område med et skeptisk perspektiv, byder det også på nye udfordringer som regulation af finansiel vejledning, bias, privatliv og sikkerhed. Lui & Lamb (2018) påpeger et behov for et passende rammeværktøj til regulering af AI i den finansielle sektor, da stakeholders i industrien i stigende grad anvender computerprogrammer og algoritmer til at rådgive deres kunder (Lui & Lamb, 2018, s. 4). AI er hverken partisk eller fordømmende, hvilket kan være en fordel for kunder i et privat og potentielt usikkert område som økonomi, og ofte også derfor, at teknologier som chatbots bliver budt velkommen af brugere på dette område. Imidlertid mangler der reguleringer i forhold til at anvende AI i bankerne, der sidder inde med sensitiv og personlig data på deres kunder, og det rejser derfor yderligere bekymringer for kundernes sikkerhed (Lui & Lamb, 2018).

Sundhed

Også i sundhedssektoren bliver AI anvendt. Anvendelsen af disse teknologier er stadig forholdsvis nyt, og det er derfor en unik mulighed for sundhedsfaglige at vejlede brugen af retfærdig, transparent og etisk AI. I sundssektoren kan AI spille en rolle i at udtrykke forskellige patientgruppers behov, og derfor er det særlig vigtigt at have strategier til at promovere retfærdige udfald (Clark et al., 2021, s. 3189). Implementering af AI bør demokratisere sundhedspleje og adressere problemstillinger relevant for socialt marginaliserede minoritetsgrupper, der oplever strukturel racisme og diskrimination i adgangen til sundhedspleje. Clark et al. (2021) bemærker, at trods der eksisterer akademisk medicinsk litteratur på området for AI-teknikker, har disse sjældent fokus på diversitet, sundhedslighed eller uligheder i sundhedsvæsenet. I tilfælde hvor disse udfordringer derimod eksplicit bliver adresseret, ses der eksempler på hvordan en sådan tilgang kan identificere og forebygge sygdomsprogression hos forskellige befolkningsgrupper. Det er vigtigt at gøre en aktiv indsats for at søge forståelse for og tilbyde løsninger til relevante behov for patienter, særligt de, der historisk har været og fortsat er marginaliseret (Clark et al., 2021, s. 3189).

Selvforstærkende bias i AI

Udover mangel på reguleringer er maskinlæring sofistikeret i sådan en grad, at selv udviklerne, der skaber algoritmerne, ikke altid ved hvordan de har udviklet sig. Det kan have konsekvenser idet, at algoritmerne kan skabe utilsigtet bias. Det kan eksempelvis opstå, hvis algoritmer der overvåger kunders tilbudsshopping, forudser at der er større risiko for misligholdelse af lån blandt disse shoppere, og disse shoppingsteder er overvejende placeret i et etnisk minioritetsområde, vil dette lede til indirekte diskrimination af denne befolkningsgruppe (Lui & Lamb, 2018). Programmører eksponerer algoritmer til store mængder data med intention om at skabe algoritmer der kan tage objektive beslutninger. Problemet opstår derfor ikke i intentionen, men når historisk data indeholder informationer om etnicitet, og algoritmen derfor analyserer mønstre mellem eksempelvis etnicitet og kriminalitet, uanfægtet om udvikleren eksplicit bad den om at lave sådan et mønster (O'Donnell, 2019, s. 548). Udvikleren vil kun vide algoritmens succesrate, og ikke hvad disse forudsigelser er baseret på. Algoritmen opdaterer sig selv konstant i takt med at den bliver eksponeret til ny data, og den originale udvikler kan derfor ikke kontrollerer denne konstante tilrettelsesproces. Som algoritmer bliver mere kontekstspecifikke, ofte baseret på millioner af faktorer, bliver det svært og ofte umuligt for udvikleren at forstå de beslutninger, algoritmen tager (O'Donnell, 2019, s. 551). Endvidere stilles følgende spørgsmål: "If algorithms

discriminate because they simply learn to discern biases that already exist in humans, why are predictive policing algorithms any worse than the status quo of biased human police?" (O'Donnell, 2019, s. 563). Problemet om forstærkende bias i maskinlæring opstår, da algoritmerne kan skabe feedback loops, hvor eksempelvis forudsigende politiarbejde bliver mere smitsom i sin etnisk relaterede skævhed over tid (O'Donnell, 2019, s. 563).

Første skridt til mere ansvarlig databrug er at udføre en kritisk undersøgelse af hvilke perspektiver, der er indkodet i dataene. Leavy et al. (2021) foreslår derfor en kritisk gennemgang af processen for datakuration, organisering og integration af data, hvor en betragtning af hvordan de involveredes identitet, perspektiver og syn på social magtstruktur kan indlejre sig i dataene og hvordan disse perspektiver stemmer overens med den tilsigtede anvendelse af det system, hvori dataene skal anvendes. Ved at kunne udpege de respektive ansvarshavende for hvert trin i udviklingsprocessen, øges transparens og ansvarlighed (Leavy et al., 2021). Muligheden for at opnå retfærdighed gennem avanceret analyse- og AI-redskaber vigtiggøre at holde os selv til ansvar for udvikling af teknologi. Proaktive strategier og innovative tilgange kan hjælpe til at reducere ulighed i hele samfundet, og løsningen er ikke nødvendigvis teknisk men menneskelig. Et samarbejde mellem dem, der bruger systemerne og dem, der møder konsekvenserne af det, eksempelvis sundheds plejepersonale og patienter, kan være et skridt mod optimale løsninger (Clark et al., 2021, s. 3191). Det er altså ikke kun tekniske løsninger kan eliminere de bias problemer, der eksisterer i AI, da disse bias er dybt indlejret i samfund. Der er behov for forståelse og sikkerhedsforanstaltninger og deraf kræves multidisciplinære tilgange for at imødekomme problemstillingerne (Ntoutsi et al., 2020, s. 10-11).

3.1.2 Etniske minoriteter

Følgende kategori er udledt af problemformuleringens andet spørgsmål, der fokuserer på problemstillinger, der relaterer sig direkte til det udvalgte bias, etniske minoriteter. Det beskriver hvorledes udfordringer opstår på baggrund af dette bias, og hvilke konsekvenser det har for brugere af denne målgruppe. Dette kan både være eksplicit, ved brugerens direkte anvendelse af teknologien, eller implicitte konsekvenser der opstår uden brugerens direkte relation til teknologien. Kategorien indeholder 21 koder fordelt på fem artikler.

AI og andre nyere teknologier byder på nye måder at realisere **ligestilling mellem etniciteter** og garantere menneskerettigheder, men det er langt fra realiteten. AI bliver anvendt til at tage beslutninger om individers liv uden menneskelig involvering, selvom den importerer mange af de bias, mennesker har. Optimister har set mulighed i, at nye teknologier kan realisere en mere ligestillet fremtid, da algoritmer udadtil kan virke til at tage objektive beslutninger baseret på datasæt. Virkeligheden er dog, at disse algoritmer reproducerer de dehumaniserende egenskaber, vi mennesker besidder, og det vækker nye udfordringer for menneskerettigheder, der søger at holde mennesket ansvarligt modsat maskiner og AI (Powell, 2018, s. 339).

AI anvendes i biomedicinsk forskning og **sundhedsrelaterede beslutninger**, og det er derfor vigtigt at udvikle upartiske AI-modeller, der fungerer ligeligt for alle etniske grupper for forebyggelse og reduktion af forskelsbehandling i sundhedssektoren. Desværre er dette ikke altid tilfældet, og uligheden i biomedicinsk data mellem etniske grupper skaber nye sundhedsmæssige uligheder gennem datadreven, algoritmebaseret forskning og kliniske beslutninger (Gao & Cui, 2020, s. 1). Biomedicinsk forskning og sundhedssektoren er i stigende grad afhængig af AI-baserede forudsigelsesanalyse til at udføre bedre diagnoser, prognoser og terapeutiske beslutninger. AI er, som tidligere afdækket, afhængig af data, og derfor er ulighed i data mellem etniske befolkningsgrupper et globalt sundhedsproblem i brugen af AI. Nyere statistikker viser, at data i prøver fra cancer genomisk forskning var fordelt således: "[...] primarily from Caucasians (91.1%), distantly followed by Asians (5.6%), African Americans (1.7%), Hispanics (0.5%), and other populations (0.5%)" (Gao & Cui, 2020, s. 2). Datasættet var altså indhentet hovedsageligt fra individer fra europæisk herkomst, og denne mangel på etnisk diversitet har ikke ændret sig meget de seneste år. Konsekvensen af dette er, at ikke-kaukasiske, som udgør 84% af verdens population, står dårligere stillet når det kommer til brug, og efterfølgende konsekvenser, af data. Konsekvenser som lav nøjagtighed for prognoser har udfald der i yderste konsekvens kan være fatale, og derfor kan denne data ulighed mellem etniske grupper siges at være en ny generation af sundhedsforskelle, når det kommer til AI (Gao & Cui, 2020, s. 2). En alment anvendt maskinlæring i de forenede stater, brugt til at understøtte sundhedsmæssige beslutninger, blev fundet at producere beslutninger ud fra etnisk relaterede bias. Systemet bruge historisk data til at forudsige fremtidige sundhedsbehov og foreslog beslutninger derefter. Den overså derfor hvordan etnisk relaterede forskelle i sundhedsvæsenet historisk

set kunne tilskrives økonomiske forhold i stedet for faktiske sundhedsbehov, og dermed blev mønstre af historisk social ulighed fastholdt (Leavy et al., 2021, s. 696).

I **politiarbejde** ses eksempler på datadreven etnisk profilering af individer, der resulterer i en gentagelse af historisk ulighed mod etniske minoritetsgrupper. Denne form for kategorisering af grupper og dertil forskelsbehandling er en fundamental underminering af retten til privatliv og ikke at blive forskelsbehandlet, og det forudsætter derfor strategier der adresserer disse problematikker. Når forudsigende algoritmer er trænet på historiske politirapporter, der i mange tilfælde er blevet kritiseret for at have åbenlyse uoverensstemmelser mellem den reelle forekomst af kriminalitet, og hvordan den er registreret. Algoritmerne bruger altså data, der i mange tilfælde er ukorrekte, og risikerer at skabe en forstærkning af eksisterende diskrimination og bias (Leavy et al., 2021, s. 696).

En forudsætning for at en algoritme er fair, er at den er kalibreret. Det betyder, at andelen af mennesker, der eksempelvis klassificeres som positive i en befolkning, skal være lig med andelen af mennesker, der er positive i hver undergruppe af befolkningen. Datasættet skal med andre ord være repræsentativt. Et eksempel hvor dette ikke er tilfældet viste følgende:

> [...] 2174/3615 = 60% of African Americans in COMPAS are considered higher risk, while this only 854/2464 = 35% for Whites; note that 1901/3615 = 53% of African Americans reoffended while this was 966/2454 = 39% for White, suggesting that African Americans actually reoffended less than predicted while the opposite is true for Whites. (Leavy et al., 2021, s. 696).

Der blev forudset højere risiko hos såkaldte "afroamerikanske" personer, men færre personer end det var forudsagt fra denne kategori gentog kriminalitet. Omvendt blev der forudset lavere risiko hos hvide, hvor flere mennesker end denne forudsigelse gentog kriminalitet. Forventningerne og de faktiske resultater stemte altså langt fra overens, og eksemplet tyder derfor på en stærk skævvridning af data og etnisk relateret bias.

Ansigtsgenkendelse har ligeledes vist at være påvirket af bias gennem datasæt. Teknologien var fundet at være mindre nøjagtig for personer med mørkere hud samt kvinder. På trods af demonstrerede etnisk bias i disse applikationer, bliver teknologier som ansigtsgenkendelse i stigende grad integreret med sikkerhed og grænsekontrol (Leavy et al., 2021, s. 696).

Internettet er et marked domineret af rigere, højere uddannede dele af verden, og derfor unægteligt fyldt med bias. Idéer og påvirkninger af lavere uddannede og af lavere indkomst er ikke på samme måde repræsenteret internettets verden, og det ses i noget så simpelt som Google søgninger. I en undersøgelse viste det, at søgninger på navne der "lyder sort" har 25% større chance for at få Google til at vise forbindelser til strafferegistre end navne, der "lyder hvide", selv når personen forbundet med navnet ikke har nogen straffeattest. Dette er blot et resultat af mange undersøgelser, der afslører Googles algoritmer, og hvor disse lærer af billioner af data, er disse unægteligt baseret på racistiske inputs: "[...] the web's values "reflect its builders—mostly white, Western men—and do not represent minorities and women [...]" (O'Donnell, 2019, s. 557).

Algoritmer kan være biased mod etniske minoritetsgrupper, selv i tilfælde hvor teknologien ikke er kodet til at tage højde for etnicitet. Et fremhævet eksempel på dette er Microsofts "Tay" eksperiment, en AI lanceret på Twitter der skulle lære af "hendes" samtaler og blive klogere med tiden. På Twitter blev botten imidlertid udsat for racistiske kommentarer og bias, der eksisterer på sociale platforme. Disse racistiske bias blev en del af Tays datagrundlag, idet Tay integrerede disse til de algoritmiske processer, og "hendes" Tweets, opslag, blev hurtigt til hadefulde angreb der blandt andet involverede opbakning til koncentrationslejre og etnisk udrensning. Botten havde internaliseret menneskelig racisme, og er blot et eksempel på maskinlærings formbare natur (O'Donnell, 2019, s. 558-559).

Flere eksempler på etnisk relateret bias illustrerer et problem i teknologisektoren. Efterfølgende afsnit vil sætte mere fokus på denne problemstilling.

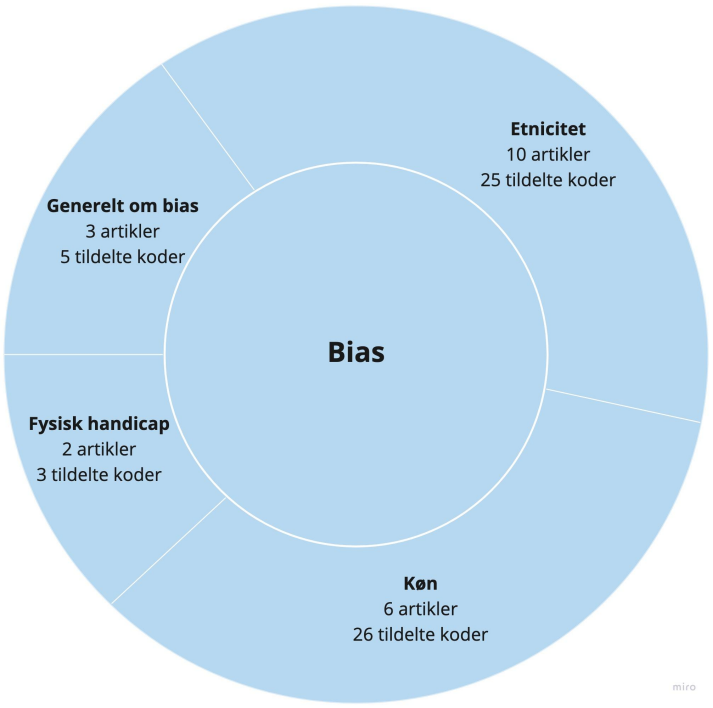
3.2 Analyse af undersøgende materiale

Følgende afsnit er en analyse af det undersøgende materiale. Artiklerne bruges særligt til at besvare det løsningsorienterede spørgsmål om hvorvidt designteorier belyst tidligere kunne implementeres for bedre udfald eller hvordan de relaterer sig til litteraturen. Afsnittet er opbygget efter de fem hovedkategorier (bilag 3), hvor hver kategori afdækker et overordnet emne og tilhørende underkategorier. Løbende illustrationer af fordelinger af koder i de forskellige hovedkategorier vil ligeledes blive præsenteret, for at skabe overblik kodningsfordeling og indhold. Her ses fordelingen mellem det fulde antal artikler anvendt i analysen, inklusiv artikler fra det problemidentificerende litteratur, og heraf hvor mange artikler der er inkluderet i hver kategori. Samme artikel kan fremgå i flere underkategorier.

Den første kategori, *Bias,* består af en gennemgang af hvilke bias problematikker, der er belyst i litteraturen samt eksempler på disse. *Problemområde* og *Teknologi* består af korte opsummeringer af, hvad litteraturen indeholder relevant til kategorien. Disse er for at skabe overblik og har en mere kvantitativ funktion. De sidste to kategorier, *Problemoprindelse* og *Metode,* går dernæst i dybden med hvorfor og hvornår problematikkerne opstår samt hvilke løsningsforslag, der er anvendt. I analysen er der fokus på de bruger- eller designorienterede artikler og passager, da disse er relevante til besvarelse af problemformuleringen. Det er ligeledes her, at teorierne i højere grad anvendes til vurdering af de enkelte tilfælde.

3.2.1 Bias

Hovedkategorien Bias har til formål at afdække hvilke biasområder og tilhørende problematikker, det indhentede litteratur belyste. Nedenunder ses en model over kodefordelingen i denne kategori, som består af fire underkategorier. Etnicitet og køn er de mest udbredte emner i denne hovedkategori, hvilket illustreres i billede 3. Ordene etnicitet og køn var en del af litteratursøgningen, og resultatet er derfor også forventeligt. Antallet af forskellige biastyper er påvirket af den oprindelige søgningsstrategi, der havde fokus på etnicitet, hvorfor disse også forekommer oftere end eksempelvis bias relateret til fysisk handicap.



Billede 3. Kodefordeling i hovedkategorien "Bias" (n=19/N=33)

Generelt om bias

Følgende underkategori er beskrivende i indholdet, hvor den har til formål at oplyse om bias som koncept samt hvilke problematikker det kan medføre. Den fungerer således indledende til resterende analyse og har en redegørende funktion.

Mennesker bliver i enhver samfundsorden opdelt i kategorier baseret på pålagte og opnåede karakteristikker, hvor nogle grupper opfattes mindre værdige end andre. Undersøgelse i algoritmisk bias baserer sig på det sociologiske koncept om kategorisk ulighed, der systemisk forskelsbehandler i adgang til ressourcer og muligheder afhængig af hvilken social kategori, et individ tilhører (Biswas, Kolczynska, Rantanen & Rozenshtein, 2020, s. 98). Denne påstand bekræftes gennem hovedanalysen, hvor tilfælde af ulige behandling på baggrund af bias vil blive eksemplificeret og undersøgt. Forskning i algoritmisk bias viser eksempler på en lang række problematikker indenfor områder som politiarbejde, sundhed og finans.

Biases eksisterer indenfor teknologiudvikling i flere former. Et af dem er historisk bias i data. Data er fyldt med historie, og historie er fyldt med bias. Et eksempel på dette er hvis et udviklingsteam markerer et datasæt som værende høj risiko indenfor banklån, og personer i teamet har bias mod særlige sociale grupper, bevidst eller ubevidst, vil dette uundgåeligt medføre uretfærdige datamærkater. Denne problematik ville kunne imødekommes med XAIs formål om transparens, der søger at skabe gennemsigtighed i, hvordan beslutninger bliver taget, og dermed holder udviklingsteamet ansvarligt. Et andet er udvælgelsesbias, hvor der opstår skævhed ved at data analyseres uden et repræsentativt nok grundlag, eksempelvis ved at en bestemt gruppe individer har større tendens til at blive udvalgt i studier (Chakraborty, Majumder & Menzies, 2021, s. 4). Udvælgelsesbias kan også forekomme i reportage, hvor én observationsform anvendes mere end andre, hvilket kan lede til udvælgelsesbias. Detektionsbias er i lighed med denne, og opstår, når ét fænomen er mere observeret end et andet.

Under- eller overrepræsentation af individer er ligeledes et bevist grundlag for bias, særligt hvis de misrepræsenterede grupper i forvejen er udsat for social ulighed og diskrimination. Disse tilfælde kan føre til diskriminerende beslutningsprocedurer, da de ikke har haft et repræsentativt grundlag. Misrepræsentation i data kan lede til repetitiv diskrimination og forfordeling. Det kan opstå ved både underrepræsentation af historisk dårligt stillede grupper, eksempelvis kvinder og ikke-hvide i teknologiudvikling og billede datasæt, men også ved overrepræsentation af eksempelvis en særlig etnicitetsgruppe i anholdelser (Ntoutsi et al., 2020, s. 4). Hvis eksempelvis politiarbejde fokuserer på beboelsesområder, der har en overrepræsentation af en bestemt etnisk målgruppe, vil dette føre til mere data på pågældende befolkningsgruppe, hvilket leder til overrepræsentation uden det nødvendigvis er et udtryk for, at målgruppen begår mere kriminalitet end andre.

Med omkring 1.8 billioner websider og 4.8 billioner brugere tilknyttet til internettet er principper som repræsentation og inklusion vigtig. Hver af disse brugere har forskellige behov, evner og præferencer, der har indflydelse på hvordan de navigerer i og opfatter information. Den høje grad af diversitet gør det vigtigt at undgå bias baseret på personlige forskelligheder som køn, kultur, alder og religion, og i stedet sikre tilgængelighed, brugervenlighed og inklusion i design af websider og teknologi (Onorati & Díaz, 2021, s. 1).

At sikre inklusion, tilgængelighed og brugervenlighed i design af web og teknologi er vigtigt for at garantere at hver eneste bruger har adgang til digital information uanset brugerens profil (Onorati & Díaz, 2021, s. 7).

Etnicitet

Følgende underkategori præsenterer eksempler på etnicitetsbias i forskellige områder. Da dette var et fokuspunkt i problemformuleringen og dertil også søgningsstrategien, er underkategorien behandlet med højst antal inkluderede artikler.

Teknologi har stor indflydelse på både eksponering og minimalisering af uligheder i sundhedssektoren. AI og maskinlæring har potentialet til enten at opretholde eller reducere systematisk ulighed i sundhedsrelaterede situationer og udfald. Der har været fokus på, at AI trænet på data der reflekterer historisk racistisk ulighed vil gentage disse uligheder og influere sundhedssystemet, hvilket der løbende vil blive set eksempler på. Flere studier har fundet beviser på, at maskinlærte algoritmer anvendt i sundhedsrelaterede områder har forskellig nøjagtighed afhængig af etnicitet. Fænomenet opstår selv efter tilfælde hvor det er forsøgt at korrigere derefter og modarbejde tendensen. Litteratur på området har undersøgt potentialet for etnicitetsbias i blandt andet vurderingssystemer, hvor disse er fundet til at have forskelligt udfald på baggrund af etnicitet. Her er der fundet systematisk undervurdering af etnisk ikke-hvide patienter (Allen et al., 2020, s. 2).

Eksempel på etnicitetsbias i almindeligt anvendte vurderingssystemer findes i en reduceret nøjagtighed af MEWS (Modified Early Warning Score), som er en tidlig vejledning anvendt af læger til at bestemme og vurdere graden af en patients sygdom. Den reducerede nøjagtighed af MEWS har vist sig ved anvendelsen på en asiatisk befolkningsgruppe, der havde konsekvent lavere nøjagtighed i forhold til hvide patienter. Beviserne indikerer, at ikke-hvide patienter er genstand for ringere sundhedspleje, og afspejler en systematisk

etnicitetsrelateret ulighed i sundhedspleje og manglende identifikation af minoritetspatienter, der med stor sandsynlighed kan have behov for øjeblikkelig behandling (Allen et al., 2020, s. 6). Blandt voksne der lever med diabetes, er dem fra etniske minoritetsgrupper, heriblandt immigranter, udenlandsk fødte, flygtninge og kulturelt marginaliserede - i øget risiko for dårlige udfald for sundhed (Pham, Gamble, Hearn & Cafazzo, 2021, s. 1). Flere studier rapporterer etnisk relaterede forskelle i kontrol, prævalens, risiko for komplikationer og diabetesrelateret dødelighed: "Data from the Centers for Disease Control indicate that non-Hispanic Black people are 2.3 times more likely to die from diabetes than their non-Hispanic White counterparts [...]" (Pham et al., 2021, s. 2). I en undersøgelse bemærkedes det, at etnicitet som egenskab ikke bidrog til forudsigelsen af glykæmisk indeks, hvor meget blodsukkeret stiger, men at disse forskelle i stedet afspejler forskelle i adgang til sundhedspleje og medicin (Pham et al., 2021, s. 3). Der er altså flere eksempler på etnicitet som en afgørende faktor i sundhedsrelaterede tilfælde, hvilket indikerer at de anvendte systemer i sundhedsområdet ikke er nøjagtige i deres udfald.

Undersøgelser har ligeledes vist, at COMPAS-algoritmen producerer risikovurderinger af recidiv, der er karakteriseret ved væsentligt højere falsk positive og en del lavere falsk negative rater for sorte tiltalte sammenlignet med hvide tiltalte. Analysen var baseret på et datasæt af tidligere dømte inklusiv udvalgt sociodemografi, kriminalitetshistorie, COMPAS-score og efterfølgende to-års tilgængelig information om tilbagefald. Der blev lavet en opfølgning på vurderingerne med en tilfældig stikprøve af 1000 tiltalte fra samme datasæt. Her fandtes lignende mønstre og niveauer af nøjagtighed af forskellen mellem falsk positive og falsk negative rater for henholdsvis sorte og hvide tiltalte, hvilket viser algoritmens systematiske unøjagtighed og etnicitetsbias. Lovmæssig anvendelse af COMPAS systemet blev udfordret i 2016, hvor argumentet var retten til retfærdig rettergang. Domstolen godkendte brugen af softwaren og understregede, at brugen alene bør fokusere på recidivrisiko (Biswas et al., 2020, s. 98). Det er bemærkelsesværdigt, at problematikken ikke alene ligger i, at ikke-hvide vurderes mere kriminelle end de er, men også at hvide mennesker vurderes mindre kriminelle, end de er. Her kunne der med fordel inddrages XAI til at forstå hvad algoritmen baserer beslutningerne på og udfordre resultaterne. COMPAS-algoritmen kan i nævnte tilfælde siges at mangle oplysningsevne, hvor uoverensstemmelserne mellem mål og realitet er konsekvent fyldte for den ramte målgruppe. Der kunne ligeledes kigges på hvilken data, der anvendes til at træne algoritmen, da dette har stor indflydelse på udfaldet.

I et studie der undersøgte millioner af udlejningslister fra webstedet Craigslist, der indeholder rubrikannoncer for blandt andet job og bolig, fandt at disse var koncentreret mod samt overrepræsenterede hvidere, rigere og bedre uddannede fællesskaber. I algoritmiske modeller havde etnicitetsbias konsekvens i form af lavere repræsentation i online udlejningslister, heriblandt sorte og latinamerikanske dele af befolkningen (Boeing, 2020, s. 461). Endvidere har boligsøgende i de forfordelte områder bedre og mere relevant information online til at understøtte deres søgning, hvor boligsøgende i andre områder er oppe imod et underskud der reproducerer mønstre af ulighed i adgang til information (Boeing, 2020, s. 463). Konsekvensen er en skævhed i muligheder, informationstilgængelighed og fastholdelse af sociale uligheder i boligmarkedet.

Sociale medier bruger, som afdækket tidligere, indholdsmoderation til at skabe tryghed med troværdig information til deres brugere. Der er dog beviser på, at moderationen ikke gælder ligeligt for alle typer af brugere, hvilket leder til uretfærdig censur relateret til brugeres køn, etnicitet eller politiske orientering. Det er blandt andet vist, at det indhold, der ofte blev fjernet fra sorte brugeres profiler, ofte relaterer sig til antiracisme og oplysning af samme. Det er også vist, at særligt sorte kvinder er den gruppe, der oplever flest tilfælde med censur (Haimson, Delmonaco, Nie & Wegner, 2021, s. 1-2).

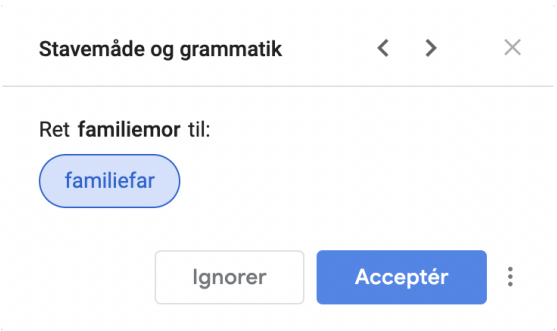
I ansigtsgenkendelsesteknologier fandt forskere, ud af effekten af etnicitet, køn og alder var præstationen i teknologien dårligst for unge, sorte kvindelige ansigter i forhold til andre grupper (Cavazos, Philips, Castillo & O'Toole, 2021, s. 1). Et studie der undersøgte tre ansigtsgenkendelsessystemer viste, at fejlraten for identificering af mørkhudede kvinders ansigter var 34,7 procent, hvoraf det for hvid mænd kun var 0,8 procent (Raz et al., 2021, s. 895). Den store forskel i præcision udgør ikke kun en risiko for minoritetsgruppen i områder som overvågning, men kan samtidig siges at underminere deres eksistens, da de ikke bliver genkendt og dermed anerkendt af udbredt dagligdags teknologi. Resultatet går imod HCDs principper om empatisk brugercentreret teknologiudvikling. XAI kan i tilfældet inddrages til blandt andet at kigge på formålet om retfærdighed, der muliggører identificering af bias i en sådan model. I politiarbejde er udfordringerne med ansigtsgenkendelse under særlig bekymring grundet algoritmiske bias. I en undersøgelse blev forskellige billeders aktiveringsværdi afprøvet og testet. Aktiveringsværdien afgør hvorvidt teknologien fungerer - des højere værdi, des bedre funktionalitet. Hypotesen, som undersøgelsen baserede sig på var, at ansigter med mørkere toner tildeles lavere værdi end ansigter med lysere toner. Testens

resultat viste, at de rekonstruerede billeder understøttede hypotesen. Aktiveringsværdien steg kun i takt med at billedet blev lysere og med hvide ansigtsegenskaber. Dette viser, at træk som hudfarve spiller en afgørende rolle i teknologiens funktionalitet, og det skaber grundlag for at adressere sådanne algoritmiske skævheder (Celis & Rao, 2019, s. 26).

Ligesom ansigtsgenkendelse har også stemmegenkendelse vist sig at være problematisk i relation til etnicitetsbias. Det er dokumenteret, at det er signifikant sværere for "afroamerikanere" at drage fordel af den stigende almene anvendelse af teknologi for stemmegenkendelse, fra virtuelle assistenter på telefoner til håndfri computerbrug for fysisk udfordrede. Denne ulighed i tilgængelighed kan aktivt skade befolkningsgruppen, eksempelvis i tilfælde hvor stemmegenkendelse bliver anvendt af ansættere til automatisk at evaluere kandidater til et job (Koenecke et al., 2020, s. 7684).

Køn

Det er allerede illustreret, at ansigtsgenkendelsesteknologier virker dårligst for kvinder, særligt med mørkere hudtoner (Celis & Rao, 2019, s. 26). Også andre teknologier viser sig at være biased mod kvinder, herunder automatiske ansættelsesværktøjer og Google Translate, verdens mest populære oversættelsesmaskine: "[...] shows gender bias. "She is an engineer, He is a nurse" is translated into Turkish and then again into English becomes "He is an engineer, She is a nurse" (Chakraborty et al., 2021, s. 1). I egen opgave blev jeg da også selv bekræftet i kønsbias' eksistens i Googles programmer, hvor Docs gav følgende forslag:



Billede 4. Eksempel på kønsbias i Google Docs, 2022.

Sådanne stereotyper er problematiske i at de vedligeholder og i nogle tilfælde forstærker bias og fastholder mennesker i forældede kønsroller.

I udvikling af servicerobotter bliver stereotyper ligeledes taget i brug, og det har en indvirkning på miljøet, den eksisterer i. Ved at tilknytte kønsnormer til robotter kan sociale kønsforskelle vedligeholdes og i nogle tilfælde forstærkes. Forskere har fundet, at specifikke arbejdsroller opfattes som enten maskulin eller feminin (Wang, Huang & Fiammetta, 2021, s. 1336-1337). Både robotters fysiske form og rolle, de bliver placeret i, kan have indflydelse på kønsnormer. Det er måske ikke så overraskende, at brugere foretrækker at blive bekræftet i normer, men det er for designere og udviklere vigtigt at være opmærksom på de konsekvenser, det kan have. Stemmeassistenter kan eksempelvis skabe skadelige kønsbias ved at forstærke forventninger til kvinders rolle i underdanige roller som assistent eller tjener. Undersøgelser har vist, at brugere foretrækker kvindelige stemmer, da disse opfattes som "varmere" (Danielescu, 2020, s. 1-2). Dette fænomen kan skabe en udfordring i at udforske kønsstereotyper i robotter, da der fra udviklerens side er en naturlig interesse i at opnå kundetilfredshed.

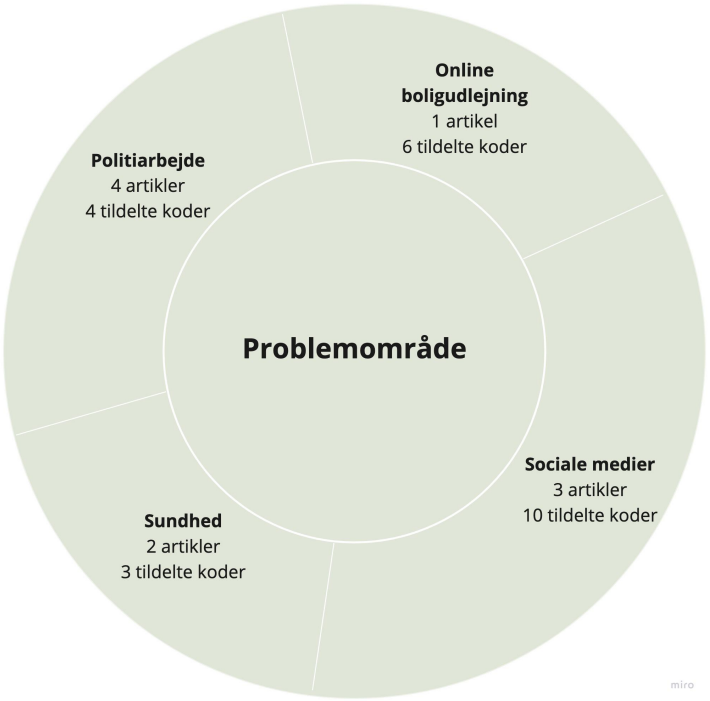
Fysisk handicap

Fysisk handicap som bias opstår, når der ikke bliver taget højde for brugere med fysiske begrænsninger, der gør teknologi utilgængelig. Det er en selvfølge at alle har lige ret til at deltage i samfundet, men realiteten i teknologiens verden er imidlertid en anden. Teknologiske værktøjer brugt i computer faget er fortsat ikke tilgængelig for personer med fysiske begrænsninger, hvilket gør det svært at inkludere disse mennesker i it-uddannelser (Luque, Veriscimo, Pereira & Filgueiras, 2014, s. 223).

Bias relaterede problematikker i teknologiudvikling er en relevant og yderst vigtigt problemstilling. At eksempelvis ansigtsgenkendelse ikke fungerer optimalt på ens mobiltelefon er ikke blot en irriterende udfordring, men det understøtter en samfundsorden der rangerer målgrupper. Set fra det perspektiv er det ikke kun et teknisk problem, men en eksistentiel problemstilling om værdi fordelt på mennesker på baggrund af ydre forhold som køn og etnicitet. Med teknologi der virker optimalt for nogle mennesker, og ikke andre, fortælles der en historie om, at en målgruppe er mere betydningsfulde end andre.

3.2.2 Problemområde

Følgende afsnit fokuserer på hvilke områder, hvori de belyste problematikker eksisterer. Afsnittet består af en kort gennemgang af de forskellige områder, hvor yderligere information om de enkelte områder vil være at finde i senere kategorier.



Billede 5. Kodefordeling i hovedkategorien "Problemområde" (n=10/N=33)

Politiarbejde

I politiarbejde er der særligt udfordringer med software til profilering som tidligere nævnte COMPAS (Biswas et al., 2020, s. 97-98). Derudover ansigtsgenkendelse til identificering af potentielle mistænkte for at øge effektivitet af undersøgelser af overvågningsvideoer (Celis & Rao, 2019, s. 25), hvilket kan være et problem når disse ikke altid fungerer optimalt og har indkodet bias.

Online boligudlejning

I online boligudlejning er der udfordringer med lige adgang til information. Fordelen ligger hos de i forvejen forfordelte befolkningsgrupper, hvilket reproducerer traditionelle uligheder og understøtter en sortering af befolkningsgrupper og adskillelsesdynamikker. Sider som

Craigslist skaber nye institutioner, der har indflydelse på dette område (Boeing, 2020, s. 449). Hvor internettet hjælper med at demokratisere informationstilgængelighed, gør det ikke dette ligeligt, når teknologiske informationsbias konstruere nye, digitale uligheder der fordrer konsekvenser som søgeomkostninger, valgmuligheder, sortering af beboere med mere (Boeing, 2020, s. 450).

Sociale medier

Mulighederne sociale medier og informationsværktøjer tilbyder er ikke lige tilgængelige for alle. Algoritmer kan gøre udvalgte personer mere eller mindre synlige via klassifikation, rangering og anbefaling. Søgemaskiner bliver i stigende grad brugt til at anbefale hvem der bør følges, citeres og ansættes. Ofte bruges disse algoritmer til at vurdere menneskers og akademiske artiklers vigtighed eller relevans, og kan dermed producere social ulighed ved at diskriminere befolkningsgrupper (Espín et al., 2022, s. 1).

Et eksempel på denne problematik er indhold eller konti der bliver fjernet, hvilket sker i højere grad for primært tre målgrupper: mørkhudede, konservative og transkønnede brugere. En undersøgelse fandt at begrundelsen for fjernelse af indhold var forskellig for de tre målgrupper. Konservatives indhold blev ofte fjernet grundet anstødeligt, upassende sprog, misinformation eller hadetaler. Transkønnede brugeres indhold blev ofte fjernet grundet kritiserende indhold af dominate målgrupper eller opslag relateret til transkønnethed. Sidst blev mørkhudede brugeres indhold ofte fjernet grundet indhold omhandlede retfærdighed og racisme-relateret indhold. De sidste to målgruppers indhold blev fjernet selvom begge tilfælde overholder sidernes reglement.

Alle tre målgrupper oplever således flest tilfælde af udelukkelse, men på forskelligt grundlag. Den konservative gruppe var den eneste hvis indhold blev fjernet grundet overtrædelse af reglement om skadeligt indhold og opretholdelse af trygge online rum med korrekte informationer. De marginaliserede gruppers indhold blev modsat fjernet i udtrykkelse af identitet og informering om retfærdighed (Haimson et al., 2021, s. 2). En af årsagerne til denne problematik er sociale mediers algoritmer, der har svært ved at skelne mellem eksempelvis racistiske og anti-racistiske opslag, da disse kan anvende samme udtryk og ordvalg i forsøg på enten at være hadefuld eller oplysende. Det resulterer i at brugere, der forsøger at belyse racisme, må bruge alias og back-up kontoer for at sikre deres indhold. At

etniske minoriteter må lave egne sikkerhedsnet blot for at kunne anvende sociale medier signalerer et behov for at adressere problemstillingen (Haimson et al., 2021, s. 7).

Sundhed

Sundhedsproblemer i relation til brug af teknologi er veldokumenteret, og er også beskrevet tidligere i opgaven. Problematikkerne indeholder blandt andet adgang til lægehjælp og kvalitet af lægehjælp for etniske minoritetsgrupper (Allen et al., 2020, s. 2).

3.2.3 Teknologi

Følgende kategori afdækker på kort vis de teknologier, der er belyst i litteraturen.

Billede 6. Kodefordeling i hovedkategorien "Teknologi" (n=14/N=33)

Kunstig intelligens

AI som teknologi er et bredt område og består således ikke at én specifik teknologisk genstand. Teknologien indgår i flere af de belyste, specifikke teknologier under denne

kategori. Teknologien udgør en underkategori da flere af artiklerne belyser AI som hovedfokus.

Algoritmisk beslutningstagen er blevet en del af vores liv i domæner som sundhed, jura, uddannelse, finans og reklame. I sundhed bruges algoritmer eksempelvis til rutinemæssigt at overvåge biokemiske signaler hos patienter hvor de advarer ved uregelmæssigheder. Algoritmer er i stand til at behandle anonymiserede elektroniske sundhedsjournaler og markere potentielle nødsituationer, som klinikere derefter har mulighed for at reagere på. På samme måde kan algoritmesystemer bruges til estimering af recidiv i politiarbejde. HR afdelinger bruger algoritmer til at filtrere jobansøgninger for at reducere arbejdstid. Universiteter anvender algoritmer til at forudse hvilke studerende der vil klare sig godt, inden de accepterer deres ansøgninger. Banker bruger mobile betalinger, og her kan betalingsservice baseret på maskinlæring verificere og identificere svindel i realtid. Forsikringsselskaber bruger automatiserede metoder til vurdering af troværdighed til hurtigt at kunne udføre komplekse godkendelser og markere usædvanlige aktiviteter. Online handelsplatforme som Amazon bruger rutinemæssigt anbefalelsesalgoritmer til at vælge hvilke varer, der er vist på brugernes sider (Shrestha & Yang, 2019, s. 1-2).

AI er en indgroet del af hverdagen og efterlader aftryk i daglige aktiviteter uanfægtet om det er ønsket eller ej. Den stigende brug er drevet af fordele som høj nøjagtighed, overlegen ydeevne, effektivitet og lave omkostninger. Derudover accelereres brugen af den massive mængde af tilgængelig data. Med flere beslutninger uddelegeret til algoritmer er der opstået beviser på etiske problematikker i relation til biases og uretfærdige beslutningsudfald. Disse udfald kan lede til skadelige konsekvenser for i forvejen minoriserede grupper (Shrestha & Yang, 2019, s. 1-2). En stigende bekymring er, at hvor AI kan ses som en objektiv måde at tage beslutninger på, opstår disse beslutninger på baggrund af data, der ikke altid er objektiv eller korrekte. Teknologien kan derfor fungere som en falsk tryghed. Det er anerkendt, at manglende transparens i beslutningsunderstøttende algoritmer og deres ofte proprietære karakter, samt deres anvendelse i sensitive område, kan lede til systematisk og institutionel diskrimination i store områder i et omfang, der kan være svært at adressere (Biswas et al., 2020, s. 97). I dette bør XAI anvendes, hvor en gennemsigtighed i beslutningernes grundlag kan opnås.

Derudover er der bekymring om den dominante rolle af et begrænset antal organisationer i en magtposition, der beslutter hvordan denne teknologi skal udvikles og anvendes. AI er en ekseptionel teknologi, der mangler regulatoriske og normative strukturer (Bélisle-Pipon, Monteferrante, Roy & Couture, 2022, s. 1-2). Hertil kan principperne fra både XAI, DT og HCD anvendes, da de søger at sikre almene borgere i udviklingsprocesser, der kan undgå en ulige magtfordeling.

Ansigtsgenkendelse

I en nyere artikel fra 2021 blev der lavet en undersøgelse, der skulle demonstrere algoritmisk bias i ansigtsgenkendelse til den almene befolkning. Baggrunden for undersøgelsen var, at denne teknologi har demonstreret at være dårligere til at genkende minoriserede målgrupper. Dette har skabt debat om lovmæssig brug af teknologien i områder som politiarbejde (Raz et al., 2021, s. 895).

Demonstrationen involverede almene befolkningsgrupper til at afprøve og teste teknologien, hvilket vækkede bekymring om dens ydeevne. Særligt vækkede det bekymring for, om teknologien blev anvendt til at determinere personers liv, som eksempelvis i rettergang. Bekymringerne kom af teknologiens usikre resultater, samt hvilke konsekvenser disse kunne have ved anvendelse i sensitive områder (Raz et al., 2021, s. 901).

Stemmegenkendelse

Stemmegenkendelse bruger sofistikeret maskinlæringsalgoritmer til at oversætte talt sprog til tekst og anvendes i virtuelle assistenter. Disse bliver anvendt til blandt andet mobiltelefoner, hjemmeapplikationer, bilsystemer, digitale diktafoner for sundhedsjournaler, automatisk oversættelse og håndfri computerbrug. Kvaliteten af teknologien er blevet bedre med årene, grundet både en forbedring af teknologien og de store datasæt, der anvendes til at træne systemerne. Bekymringen er, at teknologien ikke virker lige så godt for alle befolkningsgrupper (Koenecke et al., 2020, s. 7684). Forbedringen på baggrund af større datasæt understøtter hypotesen om, at underrepræsentation har konsekvenser for teknologiers funktionalitet.

I en undersøgelse blev nogle af de mest anvendte og anerkendte systemer for stemmegenkendelse testet, heriblandt fra Amazon, Apple, Google, Microsoft og IBM. Fordelt over alle fem systemer viste resultaterne, at antallet af fejl for mørkhudede talere var

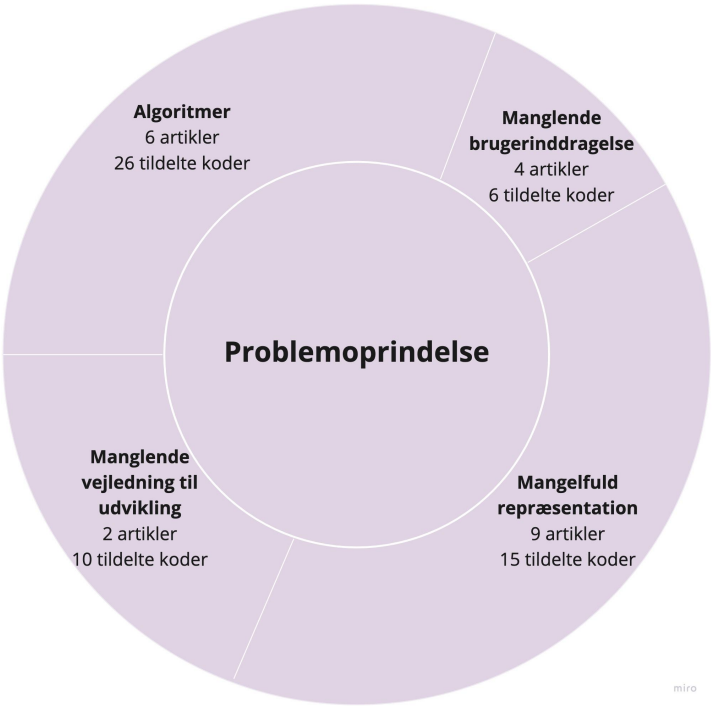
signifikant højere end for hvide talere, hvilket indikerer en skævhed i funktion for bestemte befolkningsgrupper. På trods af variationer i kvaliteten af transskriptioner i de fem systemer, var fejlraten næsten dobbelt så stor for mørkhudede talere i hver af de fem sager (Koenecke et al., 2020, s. 7685). Lignende undersøgelse viste, at disse systemer kan opnå 95% eller bedre præcision for mænd med "standard" dialekter, hvor præcisionen for kvinder er reduceret med omkring 13%. Stemmer med accent har yderligere begrænsninger med præcision på 50-80% (Danielescu, 2020, s. 1). Dette kan indikere mangel på repræsentativ datasæt, hvor algoritmen er trænet på den målgruppe, der har de bedste resultater.

Servicerobotter

Servicerobotter bliver i stigende grad anvendt for deres potentiale til at reducere omkostninger, øge effektivitet og tiltrække kunder. Robotter på frontlinjer er ofte designet med menneskelige træk og kan interagere, kommunikere og udøve service til kunder (Wang et al., 2021, s. 1336). Menneskelige robotter oplever større accept blandt kunder, og derfor bliver disse ofte anvendt til kundespecifikke formål. Måden mennesker kommunikerer på med computere kan sammenlignes med menneskelig interaktion, og af samme årsag vil brugere ubevidst tildele kønsstereotyper til robotter (Wang et al., 2021, s. 1337). Blandt andet er frisure som egenskab medbestemmende for, hvordan kønnet for en robot opfattes. Robotter med frisure viser at have tydelige kønsorienteringer, hvor kønsneutrale robotter ofte er uden frisure. Et eksperiment viste at egenskaber som frisure var i stand til at aktivere stereotypisk viden om binære køn, hvor langhårede robotter blev opfattet som mere kommunikerende og passende for stereotypisk kvindelige opgaver end de korthårede (Wang et al., 2021, s. 1339). Da robotter i stigende grad vil blive en integreret del af samfundet, er forskning på det etiske område vigtigt, da disse kan forstærke stereotyper (Wang et al., 2021, s. 1343).

3.2.4 Problemoprindelse

I følgende kategori undersøges årsager til de tidligere belyste problematikker. Fokuspunktet i denne kategori er at undersøge hvorfor problemerne opstår, og fungerer som en overgang til næste kategori, der indeholder løsningsmetoder.



Billede 7. Kodefordeling i hovedkategorien "Problemoprindelse" (n=15/N=33)

Algoritmer

Det er tidligere belyst hvordan AI og dets algoritmer kan være grund til diskriminerende teknologier samt tilhørende konsekvenser. Følgende beskriver hvordan dette fænomen opstår.

AI-baseret diskrimination kan finde sted i tre stadier:

1. Først i AI-input der er baseret på menneskelig programmering samt en given database der ligeledes er påvirket af menneskelig indflydelse. Menneskelige værdier, biases og fordomme bliver indarbejdet i programmeringen, og vil derfor blive en del af eller påvirke AI-input.

2. Dernæst AI læring der består af tre stadier: "sensing", hvor læringsalgoritmen opfanger og adopterer biased input, "reasoning", hvor denne biased database analyseres og "producing", hvor biased AI resultater produceres og integreres.
3. Sidst AI udfald, hvor AI læringsalgoritmen er udviklet baseret på menneskelige bias, bliver disse selv AI biases. De biased algoritmer kan blive en del af enhver AI applikation anvendt i den offentlige sektor, hvilket kan lede til diskriminerende AI applikationer (Weyerer & Langer, 2019, s. 509-510). Disse problematikker kan imødekommes med XAIs evaluering, hvor en evaluering af blandt andet hvilket data der er brugt, hvem der har været involveret i beslutningerne samt hvilken indvirkning resultaterne har på samfundet kan være med til at reflektere over og imødekomme risiko for bias.

Endvidere kan reproduktionen af AI diskrimination forklares via hovedsagligt tre forskellige aspekter:

1. Menneskebaseret reproduktion hvor mennesker tror på, deler og gentager diskriminatorisk indhold, hvilket endvidere anvendes som kilde for videre AI læring. Her kan XAIs formål om troværdighed med fordel inddrages, da denne søger at gøre AI forståelig og dermed mere troværdig gennem oplysning og transparens.
2. AI-baseret reproduktion hvor AI udfald anvendes som input til andre AI applikationer, hvilket kan skabe en uendelig reproduktion af fordomme og uretfærdig behandling af upriviligeret grupper. Her bør XAIs formål om overførbarhed overvejes, da det er en refleksion over hvornår det giver mening at overføre en model trænet ét sted til et andet område.
3. Forstærkning af AI diskriminerende effekt, hvor brugere opfatter teknologiske resultater til at være mere objektive og troværdige end indhold lavet af mennesker (Weyerer & Langer, 2019, s. 510). Her kan der igen drages fordel af XAIs formål om troværdighed.

Et studie fandt, at algoritmer for sprogbehandling fangede historisk diskrimination imod køn, ved eksempelvis at associere ord som "doktor" med mænd og "sygeplejerske" med kvinder. Hvor sådanne algoritmer er trænet på historisk data, vil tidligere diskrimination og stereotyper uundgåeligt reflekteres i deres resultater. Undersøgelser har ligeledes vist, at biased algoritmiske beslutninger kan opstå af forskellige årsager som menneskelige beslutninger omkring dataindsamling, beslutningstageres præferencer, udførelse af rensning

af data, valg af selve algoritmen m.fl. Disse elementer er i store træk påvirket af de valg, brugeren af algoritmen tager (Shrestha & Yang, 2019, s. 2-3).

For at kunne regulere på de bekymrende tendenser, er det vigtigt at forstå, hvordan problematikkerne opstår. I ansættelses kontekst er maskinlæringsalgoritmer trænet til at identificere egenskaber der korrelerer med stillingens kompetencebehov. Algoritmen bliver eksponeret til tidligere kandidaters ansøgninger eller videointerviews for at kunne identificere træk fra de "gode" kandidater. Algoritmen bruges dernæst til at rangere eller udsmide nye ansøgninger baseret på hvorvidt de indeholder disse træk. Algoritmen identificerer således korrelation mellem karakteristikker og udfald, og bruger disse til at forudsige udfald for fremtidige ansøgninger. Problemet opstår når ikke alle fællestræk mellem "gode" medarbejdere tilskrives kompetencer. Hvis eksempelvis de fleste "gode" kandidater er mænd, vil algoritmen forstå at maskuline egenskaber, eksempelvis brugen af maskulint sprog, er en forudsigelse for god arbejdspræstation. Algoritmen vil derfor oprangere ansøgninger, der anvender maskulint sprog, hvilket vil stille kvinder eller andre køn dårligere (Kelly, 2021, s. 902-903). Her kan med fordel anvendes XAIs globale forklaringer, der forklarer algoritmens helhed. Dette ville kunne oplyse om, hvordan algoritmen har fundet frem til dens resultater.

Manglende brugerinddragelse

En undersøgelse påpegede en demografisk undergruppes manglende repræsentation i en test distribution, hvilket resulterede i flere fejl i algoritmens ydeevne for specifikke befolkningsgrupper. Hvis eksempelvis en algoritme skulle fungere for en specifik befolkningsgruppe, måtte denne inkludere testbilleder af selv samme individer for at få algoritmen til at fungere optimalt (Cavazos et al., 2021, s. 5). Undersøgelsen indikerer manglende brugerinddragelse i test af datasæt.

I relation til stemmegenkendelsessystemer indeholder disse typisk et stort vokabular, hvilket ikke forklarer hvorfor systemerne fungerer dårligere for minoritetsgrupper. En hypotese om problemstillingen lød på, at målgruppen i den anvendte prøve brugte ord, der ikke var inkluderet i det undersøgte systems vokabular, hvilket ville forklare de oplevede forskelle i ydeevne. Hvis hypotesen er korrekt, bekræfter det blot mangel på diversitet i brugerinvolvering af det datasæt, der er anvendt til at udvikle systemet. Samme undersøgelse fandt indikation på, at de etnicitetsrelaterede forskelle i systemernes ydeevne fremkom af en forvirring af karakteristika ved såkaldt afro amerikansk-engelsk accent, hvilket sandsynligvis

opstår ved utilstrækkelig lyddata fra denne målgruppe i træningen af modellerne (Koenecke et al., 2020, s. 7686-7687). Hertil kan målsætninger fra PD om involvering af brugere med fordel anvendes, hvortil der bør tilføjes et element om krav til diversitet.

Mangelfuld repræsentation

De såkaldte scoringssystemer, der advarer om patienters helbred, viser ubalancerede patientprøver med etnicitetsrelaterede forskelle i sundhedsresultaterne. Flere undersøgelser hvor scoringssystemer er valideret viser, at resultaterne kommer fra overvejende hvide patientprøver eller i nogle tilfælde slet ikke rapporterer data fra ikke-hvide patienter (Allen et al., 2020, s. 2). Dette går imod HCDs principper om at inddrage perspektiver og have brugere - alle brugere - i centrum.

Forskelle nøjagtighed i ansigtsgenkendelse på baggrund af etnicitet er konsekvent undersøgt og rapporteret. En større undersøgelse på området viste, at erfaringsbaserede beregningsmodeller kun blev påvirket af etnicitet når grundlæggende funktioner, anvendt til at afkode ansigter, var afledt af den statistiske struktur af træningsdataene. Modellen blev som test kun trænet med enten kaukasiske ansigter som flertals etnicitet eller asiatiske ansigter som minoritets etnicitet eller omvendt. Nøjagtigheden var størst for majoriteten, uanset om denne var den ene eller anden, hvilket indikerer at nøjagtigheden for korrekt identifikation var større for den etnicitet, modellen havde størst erfaring med (Cavazos et al., 2021, s. 2). Dette bekræfter endnu engang behovet for repræsentativt datasæt i træning af sådanne modeller.

Samme problematik belyses også i teknologier om ansigtsudtryk, der er mere tilbøjelige til at opfatte aggression i en sort mands ansigt end i en hvid mands. Dette kan have flere årsager, blandt andet implicit bias i datamærkater, således at de mennesker, der lærte algoritmen at identificere følelser var mere tilbøjelige til at opfatte sorte mænd som aggressive. I XAI kaldes det for en lokal forklaring, da det forklarer forudsigelsen ved at undersøge enkelttilfælde, herunder data markeringerne. Dernæst kan problemet forklares i mangel på diversitet i træningsdata indsamlet i lande, hvor majoriteten er hvide. I denne situation lærer algoritmen at overveje en egenskab (hudfarve), der ikke reelt forudsiger følelser. Her kan der med fordel anvendes kontrastive forklaringer, der søger forståelsen for hvorfor algoritmen laver forudsigelsen for den ene målgruppe kontra den anden. Sidst, at algoritmen ikke lærer at 'læse' ansigter med mørkere hud grundet manglende eksponering (Kelly, 2021, s. 904). Her

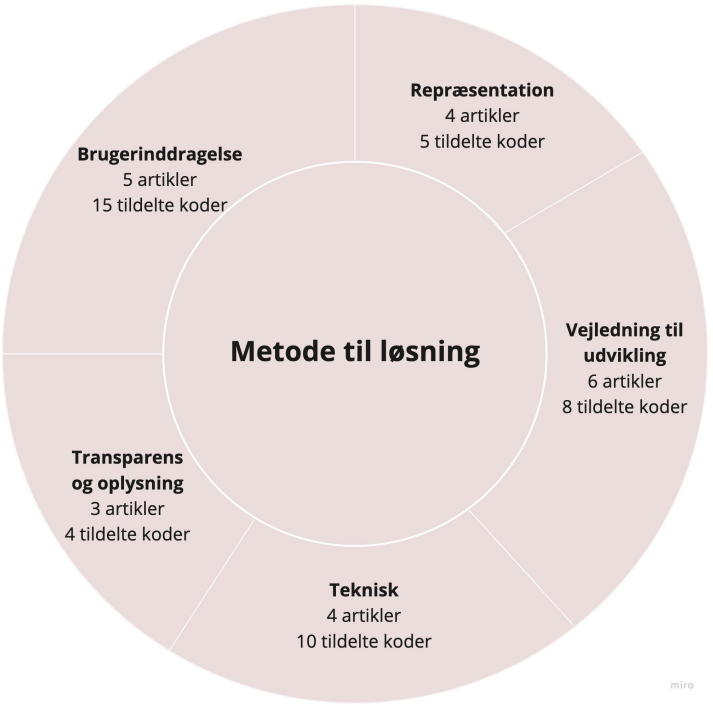
kan der anvendes kontrafaktiske forklaringer, hvor hypotetiske ændringer i input kan overvejes for bedre resultater.

Manglende vejledning til udvikling

Selvom flere etiske vejledninger er opstået de seneste år, har disse blot skabt basis for diskussion og ikke reelle reguleringer af, hvordan AI udvikles og af hvem. En undersøgelse viste, at kun 17% af virksomheder offentliggjorde den metodologi, der førte til udvikling af deres retningslinjer, hvilket skaber lav troværdighed til disse retningslinjer, da det ikke er transparent hvordan de er blevet til. Det er nemmere at øge gennemsigtigheden i AI-etiske retningslinjer end i selve AI-systemer, og dette ville være til stor gavn for troværdigheden af begge (Bélisle-Pipon et al., 2022, s. 9). XAIs formål understøtter denne opfordring, hvor transparens ville understøtte forståelsen af beslutningstagning og dermed øge troværdigheden for teknologien. Samme undersøgelse viste ligeledes, at virksomhederne der ikke involverede brugere eller eksterne interesser, i sværere grad kunne dokumentere den tilgang, der førte til formuleringen af deres etiske indsigter. Det indikerer, at disse dokumenter udarbejdes gennem en intern proces, da de ikke nævner eksterne eksperter eller andre bidragsydere. Desuden viser undersøgelsen, at brugerinvolvering ikke er steget med årene, hvilket viser at der fortsat er mangel på eksterne indsigter i udviklingen af etiske retningslinjer (Bélisle-Pipon et al., 2022, s. 7-9). Til ovenstående kan XAI være til stor gavn for virksomheder til at udvikle retningslinjer for troværdig AI.

3.2.5 Metode til løsning

I følgende og sidste kategori afdækkes de metoder, litteraturen har belyst til enten at løse problematikkerne eller som forslag til løsning.



Billede 8. Kodefordeling i hovedkategorien "Metode til løsning" (n=17/N=33)

Brugerinddragelse

Som AI etik flytter sig fra at være generelle principper til mere anvendelige, inkluderende og praktiske vejledninger skal stakeholder engagement og borgerinvolvering indlejres i en sådan rammesætning af etiske og sociale forventninger for teknologi. Etisk AI stræber efter humanistiske interesser, og det kræver en demokratisk tilgang til håndtering af sådanne teknologier, hvor de der vil blive påvirket af disse er involveret i udviklingen. Ved at involvere stakeholders, herunder brugere, bliver det muligt at afbalancere private sektorers egne interesser. Det ses derfor relevant at have diversitet i stakeholders fra flere dele af samfundet i udviklingen af regulatoriske værktøj som etiske guidelines (Bélisle-Pipon et al., 2022, s. 2). Her kan med fordel anvendes PD, der tilbyder retningslinjer for, hvordan brugerinvolvering kan realiseres. Designprocessens etiske funktion kan understøtte ønsket om

ansvarlig udvikling af teknologi, da den tager ansvar for design med hensyn til den verden, den skal placeres i.

En undersøgelse der også tidligere er nævnt, analyserede hvordan etiske guidelines udvikles, herunder mængden af involvering af stakeholders. Den viste, at lidt over en tredjedel indikerede involvering af stakeholders eksterne til organisationen, og at ingen af dokumenterne fra den private sektor involverede stakeholders (Bélisle-Pipon et al., 2022, s. 4). Industrirepræsentanter og akademikere var den mest anvendte type af engagerede stakeholders, hvor almene borgere var den mindst almindelige (Bélisle-Pipon et al., 2022, s. 5). De initiativer der involverede stakeholders, havde mange detaljer og praktikker, hvilket indikerer, at en sådan tilgang skaber mere omfattende og praktiske etiske guidelines. Endvidere påpeger undersøgelsen, at involvering af stakeholders bør være mere prioriteret, og foreslår et krav om begrundelse for tilfælde, hvor der vælges ikke at involvere flere synspunkter og interesser. På den måde bliver det ikke et krav at have involverede med i alle udviklingsprocesser, men det vil kræve en vurdering af, hvornår en sådan er nødvendig (Bélisle-Pipon et al., 2022, s. 11). I lyset af de problemer belyst gennem analysen kan det ses nødvendigt at have konkrete retningslinjer til udvikling af teknologi, hvor det er interessant at brugerinvolvering indikerer større praktik i sådanne. PD viser forøget kvalitet i produkter ved involvering af flere stakeholders i designprocesser, hvilket endvidere understøtter påstanden.

Problematikken på sociale medier, hvor marginaliserede grupper oftere får fjernet indhold eller konti, kan ligeledes imødekommes med brugerinddragelse. Der blev eksempelvis i en artikel foreslået en jury baseret tilgang, hvor en bred målgruppe af mennesker skulle moderere indhold på sociale medier. Ligeledes kunne inkludering af marginaliserede befolkningsgrupper i udviklingen af moderation politik afhjælpe problematikken. Det kræver arbejde at komme udenom gråzonerne, som lige nu gør det svært for minoriteter at tage ligeværdig del i sociale medier, og det kræver borgerinvolvering. For at lære at reducere skævheden i indholdsmoderation mellem forskellige befolkningsgrupper, bør udformningen af sociale mediers politik involvere de marginaliserede brugere. Dette indebærer at ansætte disse som konsulenter eller til at udføre forskning for at lære om deres erfaringer og om hvordan politikken kunne udvikle sig i en mere retfærdig retning. Det kan dog være problematisk at skabe tilstrækkelig mangfoldighed i en jury, da eksempelvis transkønnede personer, som var en af de mere ramte målgrupper på dette område, består af et mindretal af

befolkningen, og der derfor kan være risiko for ikke at kunne inkludere et sådant medlem (Haimson et al., 2021, s. 25-26)

I et eksempel fra et projekt kaldet "HateTrack", forsøgte forskerne at skabe et værktøj der kunne vurdere opslag på sociale medier på hvor stor sandsynlighed de indeholdte racistisk indhold. Her anvendtes PD tilgangen, hvor teamet bestod af to deltagere fra samfundsvidenskab med ekspertise i kritisk raceteori samt to deltagere med teknisk baggrund. De to første organiserede interviews med målgrupper, der oplevede hadetaler samt repræsentanter fra samfundsmæssige organisationer. Derudover indsamlede de materiale, der af interviewpersonerne blev markeret som racistisk. De brugte de udførte interviews og materialet til at generere et kodeskema, som dernæst blev anvendt af det tekniske team til at forme en algoritme, der skulle være basis for HateTrack. Værktøjet tillod derudover manuel kodning af returnerede tilfælde af falsk identificerede genstande. Projektet lærte dem om vigtigheden i at involvere de målgrupper, der var ramte af problemstillingen, da det gjorde værktøjet skarpere i at identificere tilfælde af hadetale samt tilfælde, hvor en, hvis den stod alene, ikke-nedsættende udtalelse blev nedsættende i dens kontekst. En udfordring ved metoden var, at det i nogle tilfælde kan være en retraumatiserende opgave for personer, der har været udsat for angreb at gennemgå og blive eksponeret for samme typer af angreb. Omvendt var det for andre en helende oplevelse, hvor der derfor må inkluderes sikkerhedsforanstaltninger for at passe på deltagerne. Uanset peger undersøgelsen på, at det er en fordel at inddrage brugere, der er udsat for et problem til at løse selv samme, med anerkendelse af den byrde, det måtte være for de deltagende (Leavy et al., 2021, s. 701). PD viser sig her at være til fordel for udviklingen, da den sikrer repræsentation for relevante målgrupper og deres perspektiver. Hypotesen om at marginaliserede målgruppers udelukkelse i udvikling og design af teknologi besværliggøre tilgængeligheden for selv samme, kan siges at blive bekræftet i dette eksempel.

Relateret til problemer med stemmegenkendelse bør der investeres ressourcer i at sikre at systemerne er inkluderende. Dette inkluderer bedre dataindsamling, særligt af andet end standard varianter af eksempelvis engelsk, da det ofte er her, der er størst udfordringer (Koenecke et al., 2020, s. 7688).

En anden artikel der er fortaler for brugerinddragelse omhandler AI i sundhedssektoren. De skriver, at det er nødvendigt at inddrage perspektiver af dem, der har erfaringerne, for at

kunne få succesfulde udfald. Det betyder en samarbejdsindsats der involverer patienter og fællesskaber fra forskelligartet sociale, kulturelle og økonomiske baggrunde på en bevidst og meningsfuld måde. En metode til at gøre dette er etablering af rådgivende udvalg og patientudvalg til at engagere patienter og andre relevante stakeholders. Innovative strategier omfatter projektspecifikke konsultationer, der giver patienter og samfundsmedlemmer mulighed for at give reaktioner og feedback på specifikke tilgange foreslået af sundhedspersonale og videnskabsfolk. Yderligere strategier som den såkaldte "PROM", Patient-Reported Outcomes Measures, anvendes til at vurdere resultaterne af sundhedsydelser fra patienternes perspektiv. Denne metode sikrer at nødvendige data til implementering af AI løsninger inkorporerer succesmålinger designet af forskellige patientgrupper, og dermed sikrer det brede perspektiv (Clark et al., 2021, s. 1).

Repræsentation

I en undersøgelse blev den såkaldte "Fair-SMOTE" algoritme, der har til formål at hindre bias i maskinlæringssoftware, testet. Det blev udført med en teknik, der genererer syntetiske prøver fra minoritetsgrupper og således opnår et balanceret datasæt som derefter bruges til at træne klassificering. Tilgangen kan adopteres til flere domæner i løsning af bias problemer grundet ubalancerede datasæt, herunder ansigtsgenkendelse, hvor et ligeligt fordelt datasæt af forskellige hudfarver ville anvendes. Undersøgelsen konkluderede at Fair-SMOTE antager, at bias er rodfæstet i tidligere beslutninger omkring hvilke data der bliver indsamlet og de mærkater, der bliver tildelt til dataene. Algoritmen blev fundet til at være anvendelig for bias reducering (Chakraborty et al., 2021, s. 10). I tilfælde af at der er anvendt Supervised Learning til at mærke dataene, kunne XAIs princip om retfærdighed anvendes til at undersøge hvorvidt fordomme er reflekteret i data bias samt forklaringen om ansvarlighed og data, der kan forklare bias problemet med hvem der er involveret i de forskellige trin i maskinlæringsprocessen samt hvilken data der er anvendt.

I en anden artikel menes der ligeledes, at et repræsentativt træningssæt er nødvendigt for optimering af retfærdig AI. AI forskning og udvikling bør etablere et træningssæt, der er repræsentativt for den generelle population, rapportere fordelingen af etniciteter i træningssættet samt diskutere potentielle etnisk relaterede begrænsninger. I forbindelse med at der ses tydelige, etniske forskelligheder i diabetes biomarkører, forebyggelse og udfald vil et sådant repræsentativt datasæt med høj sandsynlighed forbedre nøjagtigheden for forudsigelsesmodeller (Pham et al., 2021, s. 2-4).

Eksemplet med stereotype servicerobotter viser ligeledes, hvordan repræsentation kan lede til færre bias. Kønsliggjorte menneskelige robotter kan, i stedet for at forstærke og videreføre stereotyper udfordre de traditionelle kønsnormer. Hvis eksempelvis robotter havde udskiftelige køn kunne det udfordre tanken om det fastgjorte køn og dermed også udfordre ledende sociale normer. Et eksempel på dette var udviklingen af en mandlig robot, der opererede i et køkken for at illustrere, at design af mandlige robotter i roller som stereotypisk er kvindelige kan ændre samfundets syn på kønnene (Wang et al., 2021, s. 1337).

Fordelene ved diversitet på arbejdspladser er veldokumenterede og stammer fra inkluderingen af flere perspektiver. Diversitet i udviklingen af teknologi kan hjælpe til at løse bias ved at forbedre træningsdata og inddrage retfærdighedskoncepter i algoritmer (Leavy, 2018, s. 14). Ligeledes bør alle involverede i beslutningsprocesser være bevidste om bias relaterede problemstillinger og effekten af deres design og antagelser. Her inddrages med fordel XAIs ansvarlighedsprincipper. Studier har vist, at repræsentation relaterede bias problemer sniger sig ind i udviklingsprocesser fordi udviklingsteamet ikke er klar over vigtigheden af at skelne mellem kategorier. Eksempelvis kan medlemmer af privilegerede samfundsgrupper ikke være klar over eksistensen af kategorier som etnicitet, fordi de ser sig selv som "blot mennesker", og en sådan ubevidst standard kræver involvering af underprivilegerede grupper, som vedvarende oplever at blive opfattet som "anderledes". Denne skævhed kan imødekommes ved at forbedre diversiteten i udviklingsteams (Ntoutsi et al., 2020, s. 10).

Vejledning til udvikling

Følgende tre dimensioner kan anvendes for AI etik i udvikling: repræsentation og diversitet af perspektiver, transparens i udviklingen af etiske guidelines og håndtering af interessekonflikter og stakeholdernes egne interesser (Bélisle-Pipon et al., 2022, s. 11).

Wang et al. præsenterer en designvejledning, der kan hjælpe robotdesignere med at identificere og udfordre kønsnormer i servicerobotter (2021). Der kan argumenteres for, at denne vejledning også kan bruges til at udfordre andre normproblematikker. Vejledningen består af følgende fire trin:

1. Identificere kønsstereotyper via en brugercentreret design tilgang, eksempelvis PD, og et multidisciplinært balanceret design team. Da kvinder eksempelvis er

underrepræsenteret i teknologi og robot design, må disses perspektiver være mere tilbøjelig til at være inkluderet. Designteamet må dernæst definere kontekstbegrænsninger, herunder industri, robottens funktionsrolle og det pågældende marked for at blive bevidste om stereotyper relateret til arbejdsområdet.

2. Spore tilbage på kønsnormer i samfundet identificeret i robotten, herunder industriel, kulturel og seksuel. Hertil må designere kigge på kønsfordelingen og kønsrelaterede problematikker i industrien, blandt andet ved at kigge på industriens historie.
3. Reflektere over kønsnormers validitet og indflydelse der eksisterer i robotter og samfund. Her reflekteres over om hvorvidt et tidligere design har ubevidst bias og oversimplificeret klassifikation (eksempelvis ved ikke at have designet bukser til kvindelige robotter) samt hvorvidt robotter har udvist diversitet i forhold til de mennesker, der er i samfundet.
4. Udfordre kønsnormer ved at udvikle kønsflydende robotter. Dette ses i tidligere eksempel med den mandlige robot, der havde rolle i et køkken (Wang et al., 2021, s. 1343).

Denne form for vejledning skaber nødvendighed for refleksion over problemstillinger, der ikke altid bliver taget højde for, og derfor ubevidst bliver reproduceret i den teknologi, der bliver udviklet.

I relation til moderation på sociale medier er der behov for forskellige moderationsteknikker for forskellige brugergrupper. På nuværende tidspunkt behandles indhold på sociale medier ud fra samme moderation, hvilket efterlader minoritetsgrupper mere skrøbelige end andre. Sider som Discord og Reddit har allerede implementeret forskellige moderationsregler for forskellige kontekster, hvilket tager højde for forskellige kontekster og publikum. Som en del af sådan en tilgang må online sider tillade marginaliserede brugere at etablere særlige online forum, hvor typer af gråzone indhold får lov at blive oppe så eksempelvis indhold med oplysning og billeder af transoperationer ikke bliver fjernet (Haimson et al., 2021, s. 25-26).

Teknisk

For at være troværdig overfor datamaterialet skal det nævnes, at der er andre løsninger, end belyst i opgaven, der eksisterer i litteraturen, men af faglige årsager ikke bliver anvendt her. Dette inkluderer udviklingsrelaterede løsningsmodeller, hvor dette er uden for nærværende opgaves faglighed. Kategorien består af ti referencer fordelt på fire artikler. Disse artikler er dog blevet anvendt i andre kategorier, hvor kun det tekniske indhold er blevet udelukket.

Transparens og oplysning

En måde at imødekomme bias problematikker på, kan ske allerede i uddannelsen. En workshop der havde til formål at integrere køns inklusion i webdesign havde følgende målsætninger:

1. Opfordring til at skabe miksede arbejdsgrupper og etablere tillidsfuld, effektiv kommunikation.
2. Individuel ansvarlighed ved at sikre at alle studerende følte sig trygge i at kunne udtrykke deres meninger frit.
3. Ansigt-til-ansigt interaktion, der involverede alle gruppemedlemmer i konstruktiv diskussion til at skabe kritisk bevidsthed i at identificere sager med diskrimination og komme op med løsninger til at undgå kønsbias.
4. Samarbejdsfærdigheder, hvor studerende blev vejledt til at organisere gruppen selv og tildele roller, hvor de skulle sikre en ligelig fordeling af design og udviklingsopgaver på tværs af køn samt løse eventuelle konflikter.
5. Gruppens proces hvor hver gruppe præsenterer deres konklusioner, inklusiv oplevelser med medlemmernes interaktion og kommunikation (Onorati & Díaz, 2021, s. 3).

En sådan workshop kan skabe en bevidsthed og oplysning om bias allerede inden de studerende kommer ud og får et medansvar i at designe teknologiske løsninger.

I relation til sociale mediers moderationspolitik, kan platforme understøtte marginaliserede grupper ved at skabe transparens og ansvarlighed, og tilbyde forklaringer og detaljer der kan hjælpe brugere med at forstå deres moderationspolitik (Haimson et al., 2021, s. 24). XAI søger at undgå sådanne Black Box tilfælde ved at øge tilgængeligheden for ikke-tekniske personer at anvende og forstå AI teknologier.

AI systemer er gennemsyret af værdier fra dem, der designer og udvikler teknologi, hvilket yderligere understøttes af at privilegerede grupper i samfundet ofte ikke har været udsat for algoritmisk diskrimination. For at opnå en selvreflekterende, kritisk praksis blandt AI-udviklere foreslås en kritisk videnskabelig tilgang baseret på dekolonial teori. Her foreslås det, at dekoloniale studier sammensat med kritisk raceteori, feminisme, jura, queerness og videnskabs- og teknologistudier kan danne grundlag for en selvrefleksiv tilgang til udvikling og implementering af AI, der anerkender magt ubalancer og dets konsekvenser (Leavy et al.,

2021, s. 697). Igen belyses det, at oplysning og uddannelse, allerede inden udvikling, vil være til fordel for teknologiske løsninger.

3.3 Resultater af analyse

Ud fra ovenstående analyse er der identificeret fire vigtige elementer, som fremhæves opsummerende i følgende afsnit. De fire elementer er løsningsorienterede, og søger at give svar på hvordan de belyste problemstillinger kan imødekommes indenfor teknologiudvikling.

3.3.1 Brugerinddragelse

> *"We want to create an environment with as many different experiences, viewpoints, and ways of looking at the world and solving problems as possible. We've found that getting the opinions and insights of people with a variety of experiences to be critical in doing good work."*
>
> *(Monteiro, 2019, s. 111)*

Flere undersøgelser belyste hvordan brugerinddragelse kan være med til at reducere bias problematikker og sikre lige tilgængelighed for alle brugere. Hertil findes der flere tilgange, blandt andre PD der søger at inddrage brugere i flere stadier af udvikling. Analysen viste endvidere, at der i udviklingen af de problematiske teknologier ofte ikke blev anvendt brugercentrerede metoder eller principper - hvert fald ikke eksplicit eller dokumenteret. Kun én artikel nævnte eksplicit PD, der anvendte tilgangen til HateTrack projektet om at sikre minioritetsgrupper fra diskrimination som hadetaler. Ved at inddrage relevante fagpersoner samt brugere, der selv oplevede problemstillingen, blev værktøjet bedre til at identificere kritiske tilfælde. Det indikerer et paradoks i form af, at selvom der eksisterer bevislige metoder til reducering af bias problematikker, bliver disse ikke anvendt tilstrækkeligt i praksis.

3.3.2 Transparens

> *"Start asking questions about what you're building. Start asking questions about who benefits from what you're building. Start asking questions about who gets hurt by what you're building."*
>
> *(Monteiro, 2019, s. 90)*

I tilfælde hvor algoritmiske beslutninger viser bias relaterede skævheder, eksempelvis når én etnicitet er forfordelt i sundhed frem for en anden, bør XAI anvendes til følgende:

> 1) at forstå hvordan algoritmerne fungerer og tage beslutninger gennem oplysning og gennemsigtighed, både for brugere og udviklere
> 2) at bringe involverede til ansvar i de beslutninger, der tages undervejs
> 3) at vurdere hvorvidt en algoritme er biased og reflektere over dens mulige konsekvenser

Bias kan opstå flere steder under udvikling af teknologi. Gennem analysen blev der iagttaget to grundlæggende former for dette: teknisk og menneskelig genereret. Den ene form, teknisk genereret bias, er når algoritmen, af den ene eller anden årsag, reproducerer og forstærker menneskelig bias. Dette kan opstå ved eksempelvis brug af et datasæt til træning, der ikke er repræsentativt, eller som indeholder historisk ulige data. Tilfælde som dette kan særligt optræde ved Unsupervised Learning, hvor dataen lærer "på egen hånd". Den anden form, menneskelig generet bias, kan opstå ved moderatorer der manuelt fjerner opslag, hvor beslutningerne uundgåeligt vil være baseret på egne bias. Denne form kan forekomme i Supervised Learning, hvor data bliver mærket ved menneskelig intervention. Uanset hvilket tilfælde er der en særlig pointe i ansvarlighed omkring, at så længe en person er en del af udviklingen, som udvikler, designer eller andet, er vedkommende medansvarlig for at sikre teknologien mod biased resultater.

I tilfælde af Black Box, hvor det kan være umuligt for mennesker at forstå og fortolke udfald, er det særlig vigtigt at forsøge sig med en kritisk tilgang i brugen af resultaterne. Det kan

virke nemt og effektivt at anvende algoritmers resultater, men det kan være problematisk og uhensigtsmæssigt, hvis ikke der er bevidsthed om de mulige konsekvenser. Som tilføjelse til dette kan DT med fordel inddrages, som opfordrer til at teste under hele udviklingsprocessen. Dette ville i mange tilfælde kunne sikre mod biased løsninger. Med bevidsthed om de problematikker, der opstår på baggrund af biased algoritmer, kan det kun opfordres at sådanne beslutninger bruges med forsigtighed og som en anvisning frem for et konkret resultat. Det kræver en kritisk vurdering af resultaterne, og en viden om hvad der bør kigge efter.

3.3.3 Retningslinjer

> *"If we agree that we work within an ethical framework, as most other professions of our caliber do, then we can elevate not just the type of work we do and how we do it, but also the society which that work ends up affecting."*
>
> *(Monteiro, 2019, s. 178)*

Retningslinjer eller reguleringer ville være fordelagtige for udviklere, da disse på forhånd kunne definere hvilke problematikker, de bør være opmærksomhed på. Litteraturen illustrerer mangel på sådanne retningslinjer eller reguleringer i teknologiudvikling, der ville kunne sikre mod bias. Det ville kræve en undersøgelse i praksis at finde ud af, hvorfor de tilgængelige metoder ikke bliver anvendt, eller hvordan de bliver anvendt. Problemet er ikke desto mindre, at teknologier designes uden reelle etiske retningslinjer eller reglementer til at undgå worst-case scenarier: "We designed social networks without any way of dealing with abuse or harassment." (Monteiro, 2019, s. 10). Hvis designerne bag sociale netværk havde forestillet sig, hvordan deres medie i værste tilfælde ville blive misbrugt, kunne dette have været grobund for udvikling af retningslinjer til, hvordan sådanne tilfælde kunne undgås eller behandles. I stedet har medierne, som vist i analysen, svært ved at skelne mellem hadetaler og antiracistisk oplysning, hvor alle ender med at få samme behandling. Det resulterer i større råderum for dem, der diskriminerer, og udelukkelse af dem, der forsøger at kæmpe imod

uligheder. Pointen er at medier og andre teknologier fungerer præcist, som de er designet, og at dette kan imødekommes med retningslinjer.

3.3.4 Repræsentation

> *"All the white boys in the room, even with the best of intentions, will only ever know what it's like to make decisions as a white boy. They will only ever have the experiences of white boys. This is true for anyone. You will design things that fit within your own experiences."*
>
> *(Monteiro, 2019, s. 72)*

En af de problemoprindelser, der fremkom mest i litteraturen, var mangel på repræsentation. Det blev tydeligt illustreret, at forskellige målgrupper havde forskellige muligheder og tilgængelighed i områder som sociale medier, sundhed og politiarbejde. Dette går imod principper fra teorierne om at skabe etisk ansvarlige systemer, der har brugeren i centrum. Ved at inkludere de marginaliserede målgrupper i områder som udvikling, reglement og moderation kan disses rettigheder sikres da deres perspektiver og erfaringer vil inkluderes i udviklingen. Inddragelse af brugere kan være problematisk hvis de ikke ved, hvordan eksempelvis en algoritme fungerer. Her kan principper fra XAI med fordel anvendes til at gøre det nemmere at inddrage ikke-tekniske stakeholders, ved at gøre teknologierne mere tilgængelige.

I forhold til hvor meget repræsentation nævnes som værende årsag til de belyste problemstillinger findes det påfaldende, at ingen af teorierne indikerer et direkte krav til repræsentation. Både DT og PD nævner brugerinddragelse uden at nævne hvilke brugere, der bør inddrages. Litteraturen belyser netop hvordan særlige befolkningsgrupper gentagende gange bliver ekskluderet, og det er derfor bemærkelsesværdigt, at det ikke er en eksplicit del af teorierne. Da projektets litteraturgrundlag er begrænset med hensyn til teorierne, skal det ikke kunne afgøres, om repræsentation nævnes i de belyste teorier i andet litteratur. Til dette bør der inddrages et bredere litteraturgrundlag. Dog kan der være en pointe i, at hvis repræsentation stod som krav i brugerinddragelse, ville dette være blevet belyst i det anvendte

litteratur. Da analysen bekræfter, at repræsentation er et problem, og også en af de større, kan det derfor siges at mangle i teorierne. Bekymringerne er, at teknologier ikke virker lige vel for alle befolkningsgrupper, hvor repræsentation i form af ikke repræsentativt datasæt, mangelfuld brugerinvolvering eller lignende er en af de mest belyste problemoprindelser. Derfor kan det siges, at ingen af teorierne er eksplicitte nok i princip om brugercentrering og inddragelse. Eksempelvis siger PD, at brugerinddragelse i udvikling og test er vigtigt for at skabe de bedste løsninger, men nævner ikke noget om, hvilke brugere der er nødvendige at inddrage. Det ville eksempelvis ikke være tilstrækkeligt at involvere 100 brugere, hvis alle havde samme udgangspunkt, perspektiv og erfaring. Når problemet er så afgørende, bør der være flere overvejelser i princippet om brugerinddragelse. Overvejelser om kvalitetssikring i både udvikling og test, kunne se sådan ud:

En vurdering af hvorvidt brugerne involveret i design, test og undersøgelser er repræsentative i form af egenskaber som sociale forhold, køn, etnicitet, mental og fysisk kapacitet. Derudover sikres det, at løsningen fungerer lige godt på alle brugere.

Kapitel 4 - Diskussion og konklusion

Dette kapitel præsenterer en diskussion af den undersøgelse og analyse, der er blevet udført. Først en diskussion af identificerede elementer fra fund i analysen, der kan anvendes til videre undersøgelse. Dernæst følger en konklusion, der opsummerer projektets hovedpointer.

4.1 Diskussion

Følgende afsnit vil kort opsummere identificerede emner, der kan diskuteres eller anvendes til videre undersøgelse. Formålet er at belyse paradokser eller andre former for undren, der er dukket op under analysen af det indhentede materiale.

Kan teknologiens gevinster eksistere uden de belyste konsekvenser?

Opgaven har illustreret hvordan teknologier som AI er to sider af samme mønt, hvor den ene byder på en bedre fremtid og den anden på reproducering af social ulighed. Teknologi er i dag så udbredt, at ingen målgrupper kan sige sig fri fra den. Det er ikke opgavens formål hverken at undervurdere eller kritisere udvikling eller anvendelse teknologi, blot at sætte fokus på de problemstillinger, der løbende er illustreret. Bias problematikker vil formentligt altid være en realitet, da menneskets natur er fordomsfuldt, og det er en stor del af den historie og dermed data, der anvendes til at udvikle de systemer, vi bruger. Dette faktum lægger op til en konstruktiv revurdering af de processer, der i dag er gældende for teknologiudvikling. Det må være op til designerne af teknologien - og hertil menes alle, involverede - at tage ansvar for de udfordringer, der opstår, og opsøge metoder og retningslinjer til forebyggelse for de belyste konsekvenser. “We understand how dangerous driving can be, so we regulate cars in order to make them safer, and we license drivers in order to make sure they know the basics.” (Monteiro, 2019, s. 193). Vi ved, hvad det har af konsekvenser, nu mangler vi bare retningslinjerne.

Skal robotter forstærke eller reducere stereotyper?

Servicerobotter har som belyst fungeret understøttende for stereotyper, ved at give sig hen til brugernes forventninger om forbindelsen mellem det ydre og den indre funktionalitet. Ved at understøtte disse forventninger, forstærkes allerede stærke stereotyper - bevidste eller ubevidste. Bør udviklere underlægge sig denne tendens eller udfordre den på bekostning af omsætning og brugertilfredshed?

Mindre datasæt, mindre besvær?

Flere eksempler tyder på, at datasæt anvendt til algoritmer og maskinlæring ikke var repræsentative, og at dette resulterede i biased udfald. Som specialets titel hentyder til vil manglende diversitet ind i maskinerne føre til manglende diversitet i udfald. Tendensen kunne tyde på, at større variation i datasæt gør det sværere at konkludere og dermed foretage sig beslutninger på baggrund af disse. Black Box fænomenet kunne understøtte den hypotese, at kompleksitet i datasæt ofte fører til biased beslutninger, og at det på trods 1) alligevel anvendes og 2) ofte anses for værende tilstrækkeligt, selvom pågældende datasæt ikke repræsenterer alle brugere.

Mangler der metoder eller handling?

I underkategorien "Metode til løsning" blev der kodet 17 artikler ud af i alt 33, hvilket kan indikere, at der mangler konkrete løsninger til de belyste problemstillinger. Det kan endvidere tyde på, at en stor del af artiklerne belyser problemstillingerne uden at have et svar på, hvordan de løses. Det problemidentificerende materiale understøtter denne hypotese, hvor materialet kun indeholder artikler, der belyser problemstillingen, og altså ikke præsenterer metoder til løsning. Fordelingen mellem problemidentificerende og undersøgende materiale var næsten ligeligt, hvor førstnævnte indeholdte 14 artikler og sidste indeholdte 19. Det er en næsten ligelig fordeling mellem artikler, der blot belyser problemstillingen og artikler, der præsenterer metoder.

Endvidere kan de løsninger, der bliver præsenteret, anses banale eller have mangel på kritisk refleksion over, hvorfor metoderne ikke tages i brug - eller hvis de gør, hvorfor problemstillingerne så fortsat er relevante. Det kan eksempelvis virke fornuftigt at opfordre til inddragelse af brugere, som set i både PD og DT, men hertil mangler tilhørende refleksioner over den utilstrækkelighed, der opleves i praksis. I undersøgelsen hvor virksomheder, der udviklede etiske retningslinjer, viste sig ikke at bruge eksterne personer i processen, tyder det på at sådanne tilgange bliver udført internt. Til videre undersøgelse kunne praksis undersøges for anvendelsen af - eller mangel på - brugercentrerede metoder.

4.2 Konklusion

Opgaven har løbende illustreret de etiske udfordringer, der eksisterer indenfor design og udvikling af teknologi ud fra et Human-Centered perspektiv. Analysen afdækkede forskellige kendetegn herunder hvilke bias, område, teknologier, oprindelser og metoder litteraturen præsenterede, og skabte derigennem en kortlægning af de problematikker, der gør sig gældende indenfor området. Det blev bekræftet, at bias relaterede problematikker i teknologiudvikling er både velundersøgt og bevist i forskning. Desuden er der ikke er noget, der indikerer, at problemerne opstår med bevidsthed eller intention om de konsekvenser, det følger. Teknologi bliver udviklet med den hensigt at forbedre menneskers liv gennem effektivisering, optimering, demokratisering af viden og flere fordele, der før lå uden for menneskelige færdigheder og kompetencer. Da problemerne alligevel opstår er et sådant fokus nødvendigt for at kunne imødekomme, reducere og forhindre de bivirkninger, der følger med fordelene.

Etnicitet som udvalgt bias kan ikke kvantitativt konkluderes at være særligt udfordret på dette område ud fra det behandlede materiale i forhold til andre biases. Dog kan en kvalitativ vurdering konkludere, at der i litteraturen er særligt gældende problematiske tendenser for personer på baggrund af etnicitet, hvor undersøgelserne viser, at teknologierne fungerer bedst for hvide etniciteter. Herunder bør der også sættes fokus på etnicitetsgrupper, der stort set ikke bliver nævnt, som asiatiske befolkningsgrupper. Det hvide narrativ besværliggøre teknologiske fordele for alle, der falder udenfor denne gruppe, hvilket endvidere understøtter en magtfordeling. Det er bevist, at der i teknologiske sammenhænge er mulighed for at udfordre disse tendenser, hvor det vil være til fordel for branchen at kigge nærmere ind i problemstillingen således at de teknologier, der bliver udviklet, får en større målgruppe af mulige brugere.

Bias i andre former har ligeledes berettigelse for et fokus. Det giver endvidere ikke mening at skelne mellem hvilken målgruppe, bias problemerne har konsekvenser for, når der skal kigges ind i en løsning. Teorier indenfor Human-Centered Design viser sig brugbare i de problemstillinger, biased teknologi medfører. Om det pågældende bias er køns-, etnicitets-, handicaprelateret eller andet, er begrundelserne for problemoprindelserne de samme, og det vurderes derfor muligt at anvende teorierne uanset hvilket bias, der er gældende.

"Data is not going away. Nor are computers–much less mathematics. Predictive models are, increasingly, the tools we will be relying on to run our institutions, deploy our resources, and manage our lives. [...] these models are constructed not just from data but from the choices we make about which data to pay attention to–and which to leave out. Those choices are not just about logistics, profits, and efficiency. They are fundamentally moral."

(O'Neil, 2016, s. 218)

Litteraturliste

Allen, A., Mataraso, S., Siefkas, A., Burdick, H., Braden, G., Dellinger, R. P., McCoy, A., Pellegrini, E., Hoffman, J., Green-Saxena, A., Barnes, G., Calvert, J., & Das, R. (2020). A Racially Unbiased, Machine Learning Approach to Prediction of Mortality: Algorithm Development Study. *JMIR Public Health and Surveillance*, *6*(4), 1. SciTech Premium Collection. doi:https://doi.org/10.2196/22400

Bannon, L. J. & Ehn, P. (2012). Design: design matters in Participatory Design. I Simonsen, J. & Robertson, T., *Routledge International Handbook of Participatory Design* (s. 37-63). Routledge.

Bélisle-Pipon, J.-C., Monteferrante, E., Roy, M.-C., & Couture, V. (2022). Artificial intelligence ethics has a black box problem. *AI and Society*. Scopus. doi:https://doi.org/10.1007/s00146-021-01380-0

Bennett, C. L., Gleason, C., Scheuerman, M. K., Bigham, J. P., Guo, A., & To, A. (2021). "It's Complicated": Negotiating Accessibility and (Mis)Representation in Image Descriptions of Race, Gender, and Disability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. doi:https://doi.org/10.1145/3411764.3445498

Biswas, A., Kolczynska, M., Rantanen, S., & Rozenshtein, P. (2020). The Role of In-Group Bias and Balanced Data: A Comparison of Human and Machine Recidivism Risk Predictions. I *Proceedings of the 3rd ACM SIGCAS Conference on Computing and*

Sustainable Societies (s. 97–104). Association for Computing Machinery. doi:https://doi-org.zorac.aub.aau.dk/10.1145/3378393.3402507

Boeing, G. (2020). Online Rental Housing Market Representation and the Digital Reproduction of Urban Inequality. *Environment and Planning A*, *52*(2), 449–468. EconLit.

Booth, A., Sutton, A. & Papaioannou, D. (2016). *Systematic Approaches to a Successful Literature Review*. (2. udgave). Sage.

Cavazos, J. G., Phillips, P. J., Castillo, C. D., & O'Toole, A. J. (2021). Accuracy Comparison across Face Recognition Algorithms: Where Are We on Measuring Race Bias? *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *3*(1), 101–111. Scopus. doi:https://doi.org/10.1109/TBIOM.2020.3027269

Celis, D., & Rao, M. (2019). Learning Facial Recognition Biases through VAE Latent Representations. *Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia*, 26–32. doi:https://doi.org/10.1145/3347447.3356752

Chakraborty, J., Majumder, S., & Menzies, T. (2021). Bias in Machine Learning Software: Why? How? What to Do? *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 429–440. doi:https://doi.org/10.1145/3468264.3468537

Clark, C. R., Wilkins, C. H., Rodriguez, J. A., Preininger, A. M., Harris, J., DesAutels Spencer, Karunakaram Hema, Rhee Kyu, Bates, D. W., & Dankwa-Mullan Irene. (2021). Health Care Equity in the Use of Advanced Analytics and Artificial Intelligence Technologies in Primary Care. *Journal of General Internal Medicine*, *36*(10), 3188–3193. Research Library; SciTech Premium Collection. doi:https://doi.org/10.1007/s11606-021-06846-x

Danielescu, A. (2020). Eschewing Gender Stereotypes in Voice Assistants to Promote Inclusion. *Proceedings of the 2nd Conference on Conversational User Interfaces*. doi:https://doi.org/10.1145/3405755.3406151

Díaz-Rodríguez, N., & Pisoni, G. (2020). *Accessible Cultural Heritage through Explainable Artificial Intelligence*. 317–324. Scopus. doi:https://doi.org/10.1145/3386392.3399276

Dorst, K. (2011). The core of 'design thinking' and its application. *Design Studies*, *32(6)*, 521-532.

Espín-Noboa Lisette, Wagner, C., Strohmaier, M., & Karimi Fariba. (2022). Inequality and inequity in network-based ranking and recommendation algorithms. *Scientific Reports (Nature Publisher Group)*, *12*(1). Publicly Available Content Database; SciTech Premium Collection. doi:https://doi.org/10.1038/s41598-022-05434-1

Gao, Y., & Cui, Y. (2020). Deep transfer learning for reducing health care disparities arising from biomedical data inequality. *Nature Communications*, *11*(1). Scopus. doi:https://doi.org/10.1038/s41467-020-18918-3

Giacomin, J. (2014). What Is Human Centred Design? The Design Journal, 17(4), 606-623. doi:10.2752/175630614X14056185480186

Goethe, O., Salehzadeh Niksirat, K., Hirskyj-Douglas, I., Sun, H., Law, E. L.-C., & Ren, X. (2019). From UX to Engagement: Connecting Theory and Practice, Addressing Ethics and Diversity. I M. Antona & C. Stephanidis (Red.), *Universal Access in Human-Computer Interaction. Theory, Methods and Tools* (Bd. 11572, s. 91–99). Springer International Publishing. doi:https://doi.org/10.1007/978-3-030-23560-4_7

Gray, H. (2013). Subject(ed) to Recognition. *American Quarterly*, *65*(4), 771-798,960. Art, Design & Architecture Collection; Research Library.

Haimson, O. L., Delmonaco, D., Nie, P., & Wegner, A. (2021). Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW2). doi:https://doi.org/10.1145/3479610

Inés Alvarez-Icaza Longoria, Bustamante-Bello, R., Ramírez-Montoya, M. S., & Molina, A. (2022). Systematic Mapping of Digital Gap and Gender, Age, Ethnicity, or Disability. *Sustainability*, *14*(3), 1297. Coronavirus Research Database; Publicly Available Content Database; SciTech Premium Collection. doi:https://doi.org/10.3390/su14031297

Kamath, U. & Liu, J. (2021). *Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning*. Springer. 1-26,303-310. doi:https://doi.org/10.1007/978-3-030-83356-5

Kelly-Lyth, A. (2021). Challenging Biased Hiring Algorithms. *Oxford Journal of Legal Studies*, *41*(4), 899–928. Scopus. https://doi.org/10.1093/ojls/gqab006

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(14), 7684-7689. Research Library; SciTech Premium Collection. doi:https://doi.org/10.1073/pnas.1915768117

Kreutzer, R. T. & Sirrenberg, M. (2020). What Is Artificial Intelligence and How to Exploit It? *Understanding Artificial Intelligence. Fundamentals, Use Cases and Methods for a Corporate AI Journey*. 1-14. Springer. doi:https://doi.org/10.1007/978-3-030-25271-7

Leavy, S., Siapera, E., & O'Sullivan, B. (2021). Ethical Data Curation for AI: An Approach Based on Feminist Epistemology and Critical Theories of Race. I *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 695–703. Association for Computing Machinery. doi:https://doi-org.zorac.aub.aau.dk/10.1145/3461702.3462598

Leavy, S. (2018). Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning. *Proceedings of the 1st International Workshop on*

Gender Equality in Software Engineering, 14–16. doi:https://doi.org/10.1145/3195570.3195580

Lui, A., & Lamb, G. W. (2018). Artificial intelligence and augmented intelligence collaboration: Regaining trust and confidence in the financial sector. *Information and Communications Technology Law*, *27*(3), 267–283. Scopus. doi:https://doi.org/10.1080/13600834.2018.1488659

Luque, L., Veriscimo, E. S., Pereira, G. C., & Filgueiras, L. V. L. (2014). *Can we work together? On the inclusion of blind people in UML model-based tasks*. 223–233. Scopus. doi:https://doi.org/10.1007/978-3-319-05095-9_20

Monteiro, M. (2019). *Ruined by Design: How Designers Destroyed the World, and What We Can Do to Fix It.* Mule Books.

Neapolitan, R. E. & Jiang, X. (2018). *Artificial Intelligence. With an Introduction to Machine Learning.* (2. udgave). Taylor & Francis Group. 1-8,94-100,331-348.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., … Staab, S. (2020). Bias in data-driven artificial intelligence systems—An

introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3). Scopus. doi:https://doi.org/10.1002/widm.1356

O'Donnell, R. M. (2019). CHALLENGING RACIST PREDICTIVE POLICING ALGORITHMS UNDER THE EQUAL PROTECTION CLAUSE. *New York University Law Review*, *94*(3), 544. Research Library.

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

Onorati, T., & Díaz, P. (2021). Integrating Gender Inclusion in Web Design Courses through Design Workshops. *Proceedings of the XXI International Conference on Human Computer Interaction*. doi:https://doi.org/10.1145/3471391.3471430

Papanek, V. (2019). *Design for the Real World: Human Ecology and Social Change*. (3. udgave). London: Thames & Hudson. 3-28,54-85,215-247.

Pham, Q., Gamble, A., Hearn, J., & Cafazzo, J. A. (2021). The Need for Ethnoracial Equity in Artificial Intelligence for Diabetes Management: Review and Recommendations. *Journal of Medical Internet Research*, *23*(2). Coronavirus Research Database; Publicly Available Content Database; Social Science Premium Collection. doi:https://doi.org/10.2196/22320

Powell, C. (2018). Race and Rights in the Digital Age. *AJIL Unbound*, *112*, 339–343. Scopus. doi:https://doi.org/10.1017/aju.2018.89

Raz, D., Bintz, C., Guetler, V., Tam, A., Katell, M., Dailey, D., Herman, B., Krafft, P. M., & Young, M. (2021). Face Mis-ID: An Interactive Pedagogical Tool Demonstrating Disparate Accuracy Rates in Facial Recognition. I *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 895–904. Association for Computing Machinery. doi:https://doi-org.zorac.aub.aau.dk/10.1145/3461702.3462627

Razzouk, R., & Shute, V. (2012). What Is Design Thinking and Why Is It Important? Review of Educational Research, 82(3), 330–348. doi:https://doi.org/10.3102/0034654312457429

Ridley, D. (2012). *The Literature Review: A Step-by-Step Guide for Students*. (2. udgave). Sage Publications Ltd.

A. Robertson, T. & Simonsen, J. (2012). Participatory Design: an introduction. I Simonsen, J. & Robertson, T., Routledge International Handbook of Participatory Design. 1-17. Routledge.

Schreier, M. (2014). Qualitative content analysis. I M. Schreier. The SAGE Handbook of Qualitative Data Analysis. 170-182. London: SAGE Publications Ltd.

Sharif, S. M., Blyth, M., Ahmed, M., Sheridan, E., Saltus, R., Yu, J., Tonkin, E., & Kirk, M. (2020). Enhancing inclusion of diverse populations in genomics: A competence framework. *Journal of Genetic Counseling*, *29*(2), 282–292. SciTech Premium Collection; Social Science Premium Collection. doi:https://doi.org/10.1002/jgc4.1263

Shrestha, Y. R., & Yang, Y. (2019). Fairness in Algorithmic Decision-Making: Applications in Multi-Winner Voting, Machine Learning, and Recommender Systems. *Algorithms*, *12*(9), 199. Publicly Available Content Database; SciTech Premium Collection. doi:https://doi.org/10.3390/a12090199

Sonne-Ragans, V. (2021). *Videnskabelighed trin for trin: manual til udarbejdelse af universitetsopgaver på human-, samfunds- og sundhedsvidenskabeligt grundlag.* København: Hans Reitzel.

Sussman, R. W. (2014). *The Myth of Race: The Troubling Persistence of an Unscientific Idea*. London, England: Harvard University Press. 1-10.

Wang, Z., Huang, J., & Fiammetta, C. (2021). Analysis of Gender Stereotypes for the Design of Service Robots: Case Study on the Chinese Catering Market. I *Designing Interactive Systems Conference 2021*. 1336–1344. Association for Computing Machinery. doi:https://doi-org.zorac.aub.aau.dk/10.1145/3461778.3462087

Weyerer, J. C., & Langer, P. F. (2019). Garbage In, Garbage Out: The Vicious Cycle of AI-Based Discrimination in the Public Sector. *Proceedings of the 20th Annual International Conference on Digital Government Research*, 509–511. doi:https://doi.org/10.1145/3325112.3328220

Web og medier

Fresco, V. (Writer). 2009. *Racial Sensitivity* [Sæson 1, episode 4]. I Fresco, V., *Better Off Ted*. USA: Garfield Grove Productions. Manuskript hentet 30/05/2022 fra https://subslikescript.com/series/Better_Off_Ted-1235547/season-1/episode-4-Racial_Sensitivity

Interaction Design Foundation, *5 Stages in the Design Thinking Process*. Besøgt 16. maj 2022, fra https://www.interaction-design.org/literature/article/5-stages-in-the-design-thinking-process