Norsk KI -etikk:

Etisk implementering av KI i norsk offentlig sektor

MASTEROPPGAVE I DIGITAL KULTUR
JENNY OLSEN GEITHUS
HUMANISTISK FAKULTET
DIKULT350
HØSTEN 2024

Sammendrag

Denne masteroppgaven undersøker hvordan vi kan sikre at etiske hensyn står sentralt i implementeringen av kunstig intelligens i norsk offentlig sektor. Kunstig intelligens er raskt på vei inn i offentlig sektor, med et mål fra regjeringen om at teknologien skal bidra til økt effektivitet, bedre tjenester og opprettholdelse av både konkurransekraft og et høyt velferdsnivå. Selv om KI har potensial til å gi betydelige fordeler, har teknologien også møtt kritikk på grunn av risikoene den innebærer for utviklere, brukere, samfunnet og menneskeheten som helhet. Disse inkluderer lav forklarbarhet, fare for bias (skjevhet) og andre etiske utfordringer. Derfor er det avgjørende at Norge prioriterer å utvikle og implementere teknologi som kommer hele samfunnet til gode. Denne oppgaven er et bidrag til den pågående og høyst relevante debatten om hvordan Norge kan sikre utvikling, implementering og bruk av etiske og bærekraftige KI-løsninger. Oppgaven benytter en kombinert kvalitativ metode, bestående av en litteraturundersøkelse som analyserer nasjonal og internasjonal forskning relevant for problemstillingen, samt intervjuer med akademiske eksperter innen feltet. Oppgaven redegjør for flere kritiske punkter som må adresseres for å ivareta etiske hensyn i implementeringen av KI i Norge: Regjeringens KI-etiske tilnærming er basert på EU-kommisjonens retningslinjer for pålitelig KI, som kan føre til at kommersielle interesser kommer foran samfunnsmessige behov. KI-etiske prinsipper er abstrakt, tvetydig, lite konkret og ivaretar ikke norske verdier, som setter spørsmål til tilnærmingen som helhet. Den største spenningen ligger i at vi er enig i at vi ønsker teknologi basert på etiske prinsipper og verdier, likevel er det uenighet om hvordan dette skal realiseres i praksis og hva «etisk» faktisk innebærer. For å ivareta etiske hensyn i møte med KI argumenterer oppgaven for fire tiltak: Vi må styrke det digitale kompetansenivået i Norge, møte utfordringer gjennom en tverrfaglig tilnærming, utvikle en mer konkretisert KI-etisk praksis, og enes om hva en etisk løsning faktisk innebærer. Disse tiltakene vil kreve betydelige ressurser, en kostnad vi må akseptere om vi ønsker en etisk implementering av KI i Norge.

Abstract

This master's thesis examines how we can ensure that ethical considerations are central to the implementation of artificial intelligence in the Norwegian public sector. AI is rapidly being integrated into this sector, with the government's goal being to leverage the technology to enhance efficiency, improve services, and maintain both competitiveness and a high level of welfare. While AI holds the potential to deliver significant benefits, it has also been criticized for the risks it poses to developers, users, society, and humanity, including low explainability, the risk of bias, and other ethical challenges. As a result, Norway must prioritize development and implementation of technology that serves the public good. This thesis contributes to the ongoing and highly relevant debate on how Norway can ensure ethical development, implementation, and use of AI technology. This thesis follows a mixed qualitative method, consisting of a literature review analyzing both national and international research on the subject, as well as interviews with academic experts in the field. This thesis highlights several key points that must be addressed to safeguard ethical considerations in the implementation of AI in Norway: The government's approach to AI ethics is based on the European Commission's guidelines for trustworthy AI, but this may lead to commercial interests being prioritized over societal needs. The ethical principles of AI are abstract, ambiguous, and lack concrete definition, which raises concerns about their alignment with Norwegian values. The primary tension lies in the fact that while there is consensus on the desire for technology to be grounded in ethical values, there is still disagreement about how this should be realized in practice and what "ethical" truly entails. To address these ethical challenges, the thesis proposes four key actions: First, we must strengthen digital competence across Norway; second, adopt an interdisciplinary approach to tackle challenges; third, develop a more concrete AI ethics practice; and fourth, reach a consensus on what constitutes an ethical solution. These measures will require significant resources, a cost we must be willing to bear if we are to ensure an ethical implementation of AI in Norway.

Forord

Å skrive en masteroppgave har vært vanskelig, kjekt, altoppslukende og tidkrevende. Nå som jeg er ferdig ønsker jeg å takke alle de som har hjulpet meg i denne prosessen.

Først og fremst vil jeg rette en stor takk til Digital Kultur for fem minnerike og lærerike studieår ved Universitetet i Bergen. Programmet har gitt meg et trygt og inspirerende læringsmiljø, preget av gode klassekamerater og engasjerte professorer og undervisere. Jeg vil spesielt takke veilederen min, Ragnhild. Det som i starten virket som en nærmest umulig oppgave, ble gjennomførbar takket være din støtte. Du har hjulpet meg med å lese mellom linjene, utfordret meg til å tenke på nye og kreative måter, og vært en uvurderlig ressurs gjennom hele skriveprosessen. Dine konstruktive tilbakemeldinger og oppmuntringer har vært til stor hjelp – tusen takk! Jeg vil også takke informantene som stilte opp til intervju og delte av sin tid og kunnskap. Deres innsikt har vært avgjørende for denne oppgaven.

Til slutt vil jeg takke familie, venner og samboer, som alle har heiet og vært tålmodig når jeg har viet mye tid til oppgaven. Jeg er heldig som har hatt så mange fine mennesker rundt meg det siste året.

Jenny Olsen Geithus
November 2024

Innholdsfortegnelse

Kapittel 1: Innledning

Dette kapittelet starter med en redegjørelse for min personlige interesse for temaet, og hvordan studieforløpet har motivert meg til å undersøke implementeringen av kunstig intelligens (KI) i Norge fra et etisk perspektiv. Deretter forklarer jeg hvorfor dette temaet er viktig, og hvordan oppgaven utgjør et betydningsfullt bidrag til feltet. Jeg vil også beskrive oppgavens struktur, problemstilling og klargjøre min tolkning av begrepet «etikk» som beskriver mitt syn på oppgavens tematikk.

1.1 Bakgrunn

Som ungdom hadde jeg et nokså naivt syn på teknologi og brukte mye tid på internett uten å noen gang stille spørsmål til teknologien. Som barn og ungdom lærte man om nettvett – som blant annet handlet om hvordan man skal oppføre seg mot andre på nett, og at man ikke skal dele bilder og sensitiv informasjon med andre mennesker. Det jeg ikke forstod da, var at man ikke bare skal være kritisk til hva man deler med andre mennesker over nett, men også hva man deler med maskinen. Jeg aksepterte brukeravtaler uten å lese de, jeg ble gledelig forbauset over hvordan ulike plattformer alltid reklamerte for ting jeg likte eller hadde sett på før, og jeg tenkte ikke over at applikasjoner skulle få lov til å spore min aktivitet. Jeg visste ikke hvordan teknologiske systemer som algoritmer fungerte. Jeg visste heller ikke at maskinen ser og lagrer hva jeg gjør på internett, og legger informasjonen inn i store datasett som jeg ikke vet hva brukes til.

Da jeg begynte å studere til bachelor i Digital Kultur fikk jeg det man kaller en «AHA-opplevelse» da jeg lærte å stille spørsmål til teknologien som er en stor del av våre liv. Spesielt pensumboken *Data Feminism* (D'Ignazio og Klein 2020), som ser teknologien gjennom en feministisk linse, viste meg hvordan algoritmer ofte feiler i å representere og inkluderer det mangfoldige samfunnet vi lever i, og hvilke problemer som oppstår når data brukes for å utvikle teknologiske systemer som ikke representerer alle. Den nye innsikten førte til at jeg fordypet meg i ulike temaer knyttet til urettferdighet som følge av automatisering, og som videre har formet temaet for denne masteroppgaven. Mitt ønske for masterprogrammet i Digital Kultur var å bruke kunnskapen jeg hadde opparbeidet meg til å

utforske et område som både er viktig og meningsfullt. Etter hvert som jeg gjorde research til oppgaven, oppdaget jeg at mye av den eksisterende forskningen innen teknologiske utfordringer var rettet mot en global kontekst. Det var imidlertid mindre forskning som tok for seg de spesifikke etiske bekymringene knyttet til kunstig intelligens innenfor en nasjonal kontekst. Dette inspirerte meg til å undersøke implementeringen av KI i norsk offentlig sektor, med særlig fokus på de etiske implikasjonene.

De siste årene har behovet for etiske vurderinger i teknologisk utvikling og bruk fått økt oppmerksomhet, mye på grunn av erkjennelsen av at teknologiske løsninger kan føre til uønskede konsekvenser både for enkeltpersoner og for samfunnet som helhet (Kazim og Koshiyama 2021; Jobin Ienca og Vayena 2019). Det er likevel uenighet om hvordan etisk teknologiutvikling best kan realiseres i praksis, og uenigheter rundt hva «etisk» faktisk innebærer, fordi det er et begrep som har ulik betydning innenfor forskjellige forskningsfelt (Siau og Wang 2020). Med en vekst i ønsket om å ta i bruk KI i norsk offentlig sektor (Kommunal- og moderniseringsdepartementet 2020), blir det derfor viktig å utforske hvordan teknologien kan implementeres og anvendes på en måte som gagner hele det norske samfunnet.

I 2020 publiserte den norske regjeringen sin nasjonale strategi for kunstig intelligens. Strategien beskriver KI som et satsingsområde for Norge. Regjeringen sier at de ønsker å utnytte teknologiens innovasjonskraft for å skape effektive og brukerrettede tjenester i offentlig sektor, og for å opprettholde konkurransekraft og et høyt velferdsnivå (Kommunal- og moderniseringsdepartementet 2020, 2-5). I 2024 fremmet Norges digitaliseringsminister et ambisiøst mål om at: "innen 2025 skal 80 prosent av offentlig sektor ta i bruk kunstig intelligens (KI)." (Tung referert i Muhaisen 2024). Dette målet har imidlertid blitt kritisert for å være urealistisk. Postdoktor i sosiologi og statsvitenskap, Lisa Reutter, påpeker at det er nødvendig å først identifisere utfordringene knyttet til KI, før man utarbeider løsninger: "Det hjelper ikke bare å ønske seg mer KI. Vi trenger nok finansiering, kompetanse og klare politiske mål " (Reutter referert i Muhaisen 2024). Reutter understreker videre at det per i dag mangler tydelige retningslinjer fra myndighetene, og at den eksisterende lovgivningen fremdeles er uklar. Hun hevder også at "det ser ut som ministeren ikke forstår kompleksiteten i det å lage KI som gagner samfunnet " (Reutter referert i Muhaisen 2024).

Selv om det er uklart hva digitaliseringsministeren mente med at 80 prosent av offentlig sektor skal bruke KI innen 2025, og hvilke KI-modeller og systemer hun refererte til, er denne debatten et godt eksempel på diskursen jeg har forsket på i denne masteroppgaven. Alt skal digitaliseres og det skal skje fort, men hvordan sikrer vi at teknologien vi utvikler gagner hele samfunnet? Denne oppgaven er et bidrag inn mot den høyaktuelle debatten om hvordan Norge kan utvikle og implementere gode og etiske løsninger for KI.

1.2 Problemstilling og oppbygning

KI-teknologi er i stor fart på vei inn i offentlig sektor (Kommunal- og moderniseringsdepartementet 2020), og som Reutter understreker er det en kompleks utfordring å lage KI som gagner hele samfunnet (Reutter referet i Muhaisen 2024). De siste årene har flere stilt seg kritisk til teknologien, fordi flere mener den utgjør en betydelig risiko for utviklere, brukere, samfunn og menneskeheten med lav forklaringsevne, fare for bias (skjevhet) og andre etiske bekymringer (Siau og Wang 2020, 74).

For å sikre at teknologi som utvikles og brukes er etisk forsvarlig, har en av de mest utbredte tilnærmingene vært å regulere teknologien gjennom å utvikle retningslinjer og KI-etiske prinsipper (Kazim og Koshiuama 2018; Sieu og Wang 2020; Hickok 2020; Jobin, Ienca og Vayane 2018). De siste årene har det blitt utgitt en rekke retningslinjer og KI-etiske prinsipper fra ulike private og offentlige bedrifter og etater (Hickok 2020). En av disse er EU kommisjonens retningslinjer for pålitelig kunstig intelligens (HLEG 2019), som er bygget på KI-etiske prinsipper. EU sine retningslinjer legger i stor grad føringer for den norske regjeringens KI-etiske tilnærming, ved at regjeringen har valgt å følge EU's KI-etiske prinsipper for å sikre pålitelig og tillitsvekkende KI i Norge (Kommunal- og moderniseringsdepartementet 2020, 58-60). KI-etiske prinsipper har blitt kritisert av flere, for å blant annet være tvetydig og abstrakt, og for å mangle politiske og sosiale kontekster som bidrar til et gap mellom teori og praksis **(**Kazim og Koshiuama 2018; Munn 2022; Jobin, Ienca og Vayane 2018). Teknologien som utvikles, brukes og implementeres i Norge, bør gagne hele samfunnet. Hvordan kan vi sikre at teknologien vi bruker er etisk forsvarlig? For å realisere dette ønsket, er det behov for vurderinger av hvordan Norge håndterer KI-etikk, og

hvordan vi kan sikre at Norges tilnærming er konkret og meningsfull, snarere enn generell og overfladisk. Disse spørsmålene legger grunnlaget for oppgavens problemstilling:

> *Hvordan kan vi sikre at etiske hensyn står sentralt i implementeringen av kunstig intelligens i norsk offentlig sektor?*

Med denne problemstillingen ønsket jeg å finne ut av hvorvidt Norges KI-etiske tilnærming ligger til rette for etisk bruk, utvikling og implementering av KI i Norge. Problemstillingen besvares gjennom en mikset kvalitativ metode som bygger på forskning og offentlige dokumenter. Oppgaven benytter en kvalitativ-mikset metode bestående av en litteraturundersøkelse på både nasjonalt og internasjonalt nivå, samt ny empirisk kunnskap gjennom semi-strukturerte intervjuer med eksperter fra akademia. Gjennom en tematisk gjennomgang av datamaterialet ble det avdekket flere utfordringer innenfor etisk vurdering: *KI-etiske prinsipper*, *internasjonal påvirkning*, og *menneskene bak systemene*, og identifisert fire sentrale områder der det er behov for videre arbeid for å realisere oppgavens mål: *tverrfaglig samarbeid* og *digital kompetanse*, *en mer konkretisert tilnærming*, og *enighet i hva en etisk løsning innebærer i praksis*.

1.3 KI-etikk som konseptuelt rammeverk

Denne masteroppgaven bygger på et KI-etisk perspektiv hentet fra litteratur fra feltet. Kunstig intelligens er et komplekst og tverrfaglig fenomen, og KI-etikk er et felt som analyserer implikasjonene av KI som omfatter flere disipliner med ulike begrepsrammer. IBM beskriver KI-etikk som: “KI-etikk er et tverrfaglig felt som studerer hvordan man kan optimere KI sin positive innvirkning samtidig som man reduserer risiko og uønskede utfall” (IBM u.å, egen oversettelse). Som nevnt tidligere har det de siste årene blitt utgitt en rekke retningslinjer og KI-etiske prinsipper for å regulere teknologien. En del av KI-etisk forskingslitteratur tar for seg disse prinsippene og retningslinjene og evaluerer deres praktiske virkning, som jeg vil gå gjennom i oppgavens litteraturundersøkelse (Kazim og Koshiyama 2021; Munn 2022; Jobin, Ienca og Vayane 2018). Som IBMs beskrivelse av KI-etikk sier, så ønsker KI-etikken å redusere risikoen for uønskede utfall. Et eksempel på slike uønskede utfall som ofte diskuteres i KI-etisk forskningslitteratur, er bias (Buolamwini og Gebru 2018; Kazim og

Koshiyama 2021; D'Ignazio og Klein 2020). Denne «grenen» av KI-etikk, tar både for seg de tekniske sidene bak bias som datasett og datamerking (Danks og London 2017; Ferrer mfl. 2021; D'Ignazio og Klein 2020), og menneskelig skjevhet - som rase og kjønns bias og hvordan de videreføres inn i algoritmiske systemer (Buolamwini og Gebru 2018; D'Ignazio og Klein 2020; Howard og Borenstein 2017; Mullaney 2021; Leavy, O'Sullivan og Siapera 2020).

Målet med oppgaven er å analysere hvorvidt Norges KI-etiske tilnærming legger til rette for etisk bruk, utvikling og implementering av KI. Derfor har jeg valgt å bruke KI-etisk litteratur, fordi det kan bidra til å belyse hvilke etiske prinsipper som understøtter Norges KI-strategi, og hvordan disse prinsippene omsettes i praksis. Med et KI-etisk perspektiv og et humanistisk verdigrunnlag som setter mennesket i sentrum, fokuserer jeg på de sidene ved KI som har betydning for etiske vurderinger. Siden oppgaven overordnet hviler på et humanistisk verdigrunnlag, vil dette verdigrunnlaget være gjennomgående, samtidig som innsikter fra andre fagområder vil trekkes inn for en mer helhetlig belysning av problemstillingen. Jeg anser ikke KI-etikk som et fast teoretisk rammeverk, men snarere som en samling av ulike teorier og perspektiver. Likevel ligger det et verdigrunnlag i KI-etisk litteratur som gjennomstyrer oppgaven, knyttet til de moralske, etiske og sosiale konsekvensene av KI-utvikling og bruk. Denne litteraturen gir også grunnlag for en kritisk vurdering av teknologien, og fungerer som et konseptuelt rammeverk for å evaluere teknologiske utfordringer i offentlig sektor.

Akkurat som «KI» selv, er «etikk» et komplekst og flertydig fenomen. Det finnes mange ulike tolkninger av hva det innebærer å handle etisk (Siau og Wang 2020). Fordi oppgavens problemstilling og tematikk handler om vurderinger av det etiske og har som mål å vurdere om Norge legger til rette for etisk bruk, utvikling og implementering av KI i norsk offentlig sektor, er det derfor viktig å klargjøre hvordan jeg som forsker forstår og anvender etikkbegrepet, da denne forståelsen danner grunnlaget for resten av oppgaven. En definisjon som sammenfatter med hvordan jeg anvender begrepet etikk i oppgaven, er Datatilsynets (2024): «'Etisk KI' refererer primært til å justere kunstig intelligente systemer i tråd med etiske prinsipper og verdier. Det kan være å sikre at systemet ikke viderefører fordommer eller urettferdighet, og at de bidrar positivt til menneskelig velferd og rettigheter.»

Kapittel 2: Teoretisk bakgrunn

I dette kapittelet vil jeg gå gjennom sentrale teorier som vil være viktig for å besvare problemstillingen. Kapittelet starter med å redegjøre for norsk KI-etisk tilnærming, med utgangspunkt i regjeringens nasjonale strategi for kunstig intelligens (Kommunal- og moderniseringsdepartementet 2020). Videre vil den teoretiske bakgrunnen beskrive framveksten av KI-etikk, som er et forskningsfelt som ser på teknologien gjennom en kritisk linse (Siau og Wang 2020), og redegjøre for det Kazim og Koshiyama (2021, 6) beskriver som en av de mest utbredte tilnærmingene for å sikre etisk KI: «retningslinjer og KI-etiske prinsipper». I de tre siste delkapitlene vil jeg se nærmere på etikk sett i lys av KI-etikk, se på grunnleggende teknologiske prosesser innenfor KI, og gjøre rede for humanioras rolle i teknologisk utvikling.

2.1 Norsk KI-etisk tilnærming

I det første delkapittelet i den teoretiske bakgrunnen vil jeg gjøre rede for regjeringens KI-etiske tilnærming ut ifra regjeringens dokument «Nasjonal strategi for kunstig intelligens» kapittel fem «Ansvarlig og pålitelig kunstig intelligens» (Kommunal- og moderniseringsdepartementet 2020) som ble utgitt i 2020. Gjennom strategien kommer det tydelig fram et ønske om at Norge skal satse stort på digitalisering og KI de kommende årene. Regjeringen vil ta i bruk KI-teknologi for å kunne opprettholde et høyt velferdsnivå i framtiden (Kommunal- og moderniseringsdepartementet 2020, 2). Strategien peker på hvordan KI vil kunne bringe store muligheter for både enkeltmennesker og for samfunnet, og at teknologien vil kunne bidra til effektivisering innen næringsliv og offentlige tjenester (Kommunal- og moderniseringsdepartementet 2020, 5). Regjeringen ønsker at Norge skal utnytte innovasjonskraften til KI, og ta en ledende posisjon i å anvende KI innenfor helse, olje og gass, energi, maritim, næring og offentlig sektor (Kommunal- og moderniseringsdepartementet 2020, 7).

2.1.1 Tillitt

Regjeringen skriver innledningsvis i kapittel fem at *tillit* er et viktig prinsipp i Norge, og mener at «Norge er kjennetegnet av at vi har høy tillit til hverandre og til statlige og private virksomheter» (Kommunal og Moderniseringsdepartementet 2020, 56). Derfor har regjeringen som mål å opprettholde og forsterke tillit gjennom implementeringen av KI. Regjeringen presenterer fem punkter for å ivareta dette tillitsforholdet: 1) «kunstig intelligens som utvikles og brukes i Norge skal bygge på etiske prinsipper, og respektere menneskerettighetene og demokratiet», 2) «forskning, utvikling og bruk av kunstig intelligens i Norge skal bidra til ansvarlig og pålitelig kunstig intelligens», 3) «utvikling og bruk av kunstig intelligens i Norge skal ivareta den enkeltes integritet og personvern», 4) «digital sikkerhet skal bygges inn i utvikling, drift og forvaltning av løsninger som bruker kunstig intelligens», og 5) «tilsynsmyndigheter skal føre kontroll med at systemer basert på kunstig intelligens på sitt tilsynsområde opererer innenfor prinsippene for ansvarlig og pålitelig bruk av kunstig intelligens» (Kommunal og Moderniseringsdepartementet 2020, 56-57).

2.1.2 Spørsmål reist av regjeringen

Videre i kapittel fem i nasjonal strategi for KI redegjør regjeringen for viktige problemstillinger i møte med KI, hvor de mener at utvikling og bruk av KI kan skape utfordringer og reise vanskelige spørsmål, spesielt ved KI-modeller som bygger på personopplysninger (Kommunal- og moderniseringsdepartementet 2020, 57). Nedenfor vil jeg gjenfortelle regjeringens punkter, fordi de er viktig for diskusjonen senere i oppgaven.

Personvern: Regjeringen (Kommunal- og moderniseringsdepartementet 2020, 57) skriver at det kan bli utfordrende å ivareta personvern når man tar i bruk en teknologi som trenger mye data når den utvikles. I Norge har vi det regjerningen kaller et «grunnleggende personvernsprinsipp om dataminimering», dette innebærer at vi skal begrense mengden innsamlede personopplysninger til det som er nødvendig for å oppfylle formålet med innsamlingen. På denne måten kan teknologiens behov for store datamengder stride mot dataminimering (57).

Datakvalitet: En annen utfordring regjeringen redegjør for, er fare for bias (Kommunal- og moderniseringsdepartementet 2020, 57). For å minimere risiko for bias, mener regjeringen at det vil kreve data med «god kvalitet og struktur» (57). Derfor mener regjeringen at hver enkel virksomhet må ha oversikt over hva datagrunnlaget de bruker betyr, brukes til, hvilke prosesser de inngår i, og om det finnes rettslig grunnlag for å dele dataene. Regjeringen reiser også bekymring til hvordan datasett brukt for veiledet læring, som jeg vil se nærmere på i delkapittel 2.3.2, fordi de kan inneholde skjevheter fra «menneskelige feilvurderinger eller historiske skjevheter» (s, 57) i datagrunnlaget. Regjeringen beskriver også at KI kan bli påvirket ut ifra hvem som definerer problemstillinger (58).

Mangel på transparens: Regjeringen (Kommunal- og moderniseringsdepartementet 2020, 58) skriver at det KI-etiske prinsippet *transparens,* kan bli utfordrende å ivareta når dyplæringsalgoritmer er en «sort boks», som gjør det vanskelig med innsyn i hvordan modellen fatter beslutninger. På områder hvor det er viktig å kunne forklare hvordan algoritmiske beslutninger fattes, skriver regjeringen at det kan være et alternativ å velge andre tilnærminger enn dyplæring (58).

Autonomi: Den siste bekymringen reist av regjeringen er knyttet til distribusjon av ansvar når noe går galt (Kommunal- og moderniseringsdepartementet 2020, 58). Kunstig intelligens opererer med en viss grad av autonomi, noe som gjør at systemene kan ta beslutninger og fatte beslutninger uten menneskelig involvering. Selv om graden av autonomi kan variere, understreker regjeringen at det reiser spørsmål om hvem som skal holdes ansvarlig for konsekvensene av beslutninger, og hvordan en slik autonomi bør avgrenses (58).

2.1.3 Regjeringens prinsippbaserte tilnærming

I 2019 publiserte EU-kommisjonens ekspertgruppe etiske retningslinjer for pålitelig bruk av KI (HLEG 2019). Retningslinjene er basert på grunnleggende rettigheter og internasjonal menneskerettighetslovgivning. Regjeringen skriver at formålet med retningslinjene er: «å fremme ansvarlig og bærekraftig utvikling og bruk av kunstig intelligens i Europa» (Kommunal- og moderniserings departementet 2020, 57). EU-kommisjonens ekspertgruppe mener at KI må være lovlig, etisk og sikker for å kunne defineres som pålitelig og tillitsvekkende. Med utgangspunkt i dette har HLEG (2019) foreslått syv prinsipper for etisk

og ansvarlig utvikling av KI. Regjeringen tar utgangspunkt i disse prinsippene for ansvarlig bruk og utvikling av KI i Norge (Kommunal- og moderniseringsdepartementet 2020, 58). Prinsippene sier at: 1) «KI-baserte løsninger skal respektere menneskets selvbestemmelse og kontroll», 2) «KI-baserte systemer skal være sikre og teknisk robuste». 3) «KI skal ta hensyn til personvernet», 4) «KI-baserte systemer må være gjennomsiktige», 5) «KI-systemer skal legge til rette for inkludering, mangfold og likebehandling», 6) «KI skal være nyttig for samfunn og miljø», og 7) «Ansvarlighet» (Kommunal- og Moderniseringsdepartementet 2020, 59-60).

Regjeringen skriver at disse prinsippene i hovedsak er rettet mot KI som bygger på data fra mennesker eller som har innvirkning på mennesker, og for industriell bruk av KI som ikke bruker personversopplysninger (Kommunal- og Moderniseringsdepartementet 2020, 59). Ifølge regjeringen kan det være vanskelig å oppfylle alle syv prinsipper samtidig, fordi det kan oppstå spenninger mellom dem. Da kan det bli nødvendig å foreta det regjeringen beskriver som avveiinger, som må gjøres på en rasjonell og metodisk måte. Regjeringen skriver: «hvis det ikke er mulig å identifisere en etisk akseptabel avveiing mellom de ulike kravene, bør utviklingen, utbredelsen og anvendelsen av den aktuelle løsningen ikke fortsette i den samme formen» (59). Videre understreker regjeringen at alle avveiningsbeslutninger skal være godt begrunnet og dokumentert. Dersom en KI-basert løsning medfører urimelige negative konsekvenser, bør det være etablert mekanismer som gjør det mulig å rapportere slike virkninger (59).

Innebygd personvern og etikk: Regjeringen skriver at algoritmer kan kontrolleres gjennom innsyn og revisjon, og beskriver det som «hensiktsmessig» for brukere og utviklere å integrere etiske hensyn og personvern fra start når man utvikler KI-systemer (Kommunal- og Moderniseringsdepartementet 2020, 60). Regjeringen påpeker at dette er en tilnærming som allerede er godt etablert innen personvernsfeltet, der innebygd personvern er et krav i personvernsforordningen, og at personvernet skal ivaretas i alle faser av utviklingen. Dette er nødvendig fordi informasjonssystemer må oppfylle forordningens krav og beskytte individers rettigheter. Regjeringen mener at etiske vurderinger skal på tilsvarende måte integreres i algoritmeutvikling, som vil innebære vurdering av om algoritmen er sårbar for manipulasjon eller bidrar til diskriminering. Etiske vurderinger kan også omfatte vurderinger av

algoritmenes miljøpåvirkning, og i hvilken grad løsningen fremmer FN's bærekraftsmål, skriver regjeringen. For at innebygd etikk og personvern skal kunne realiseres, mener regjeringen at de som utvikler KI-baserte løsninger må enten ha eller tilnærme seg nødvendig kompetanse og henviser til høyere utdanningsinstitusjoner som bør vurdere hvordan etikk og personvern kan integreres i fag som datavitenskap og informatikk (60).

Utfordringer for forbrukere: Regjeringen mener at selv om KI kan forenkle hverdagen til forbrukere, kan det også oppstå bekymringer knyttet til personvern, transparens og forbrukerrettigheter (Kommunal- og Moderniseringsdepartementet 2020, 61). Regjeringen beskriver forbrukere som spesielt sårbare når KI brukes til å utvikle personaliserte tjenester og målrettet markedsføring, som er basert på innsamling og behandling av deres personopplysninger. På et internasjonalt nivå har en økende bekymring vært at bedrifter ikke ivaretar forbrukernes personvern godt nok (61). En undersøkelse fra Consumers International (Consumer International 2019, referert i kommunal- og moderniseringsdepartementet 2020 s, 61) viste at forbrukere verdsetter blant annet økt selvstendighet som er en virkning av automatisering, men at de også var usikre på hvordan opplysninger brukes og hvem som behandler dataene, og etterlyste mer kontroll og større tydelighet (61).

Når digitale tjenester og markedsføring i stadig større grad individualiseres, mener regjeringen at forbrukere risikerer å bli utsatt for forskjellsbehandling og «vilkårlige beslutninger uten transparens» (Kommunal- og Moderniseringsdepartementet 2020, 61). Ifølge regjeringen kan dette kan bidra til manipulasjon og at forbrukere blir påvirket til å ta valg de vanligvis ikke hadde tatt, og at KI påvirker forbrukeres sosiale liv på mange måter, og ulike sektorer i samfunnet. Regjeringen skriver at bruken av KI reiser juridiske spørsmål og at bruken av KI gir opphav til juridiske utfordringer innenfor flere lovområder, særlig knyttet til konkurranse-, personvern- og forbrukervernlovgivning. Det er derfor viktig at relevante tilsynsmyndigheter samarbeider på dette feltet. Dette innebærer å utveksle kunnskap og informasjon, samt deltakelse i internasjonale fora (56).

2.2 KI-etikk: Regulering og prinsipper

Kunstig intelligens har stor betydning og innvirkning på fremtiden, særlig innen medisin og helse, arbeidsmarkedet, velferd og samfunnsstyring. Bruk av KI kan ha positive virkninger ved å redusere menneskelig feil og skjevheter, og opprettholde et høyt sikkerhetsnivå når teknologien brukes «riktig» (Siau og Wang 2020, 74). Samtidig har det i de siste årene blitt avdekket hvordan KI-teknologi implementeres uten tilstrekkelig forståelse av de sosiale skjevhetene som er innebygd i systemene, eller av de samfunnsmessige spørsmålene teknologien reiser (Frank mfl. 2019, 79). KI-teknologi kommer i mange former og anvendes i stor skala. Etter hvert som KI blir mer utbredt mener Frank, Wang, Cebrian og Rahwan (2019, 79) at forskere og beslutningstakere må balansere de negative og positive implikasjonene ved teknologien.

KI teknologi er komplekst, fordi teknologien kan gi klare fordeler, samtidig som den reiser spørsmål i om den er etisk forsvarlig. Siau og Wang (2020, 74) mener at KI-teknologi har oppnådd mye bra fra et teknisk perspektiv, for eksempel med teknologier som ansiktsgjenkjenning og selvkjørende biler. I følge Siau og Wang har KI gitt store fordeler, som økonomisk vekst, sosial utvikling, forbedring av menneskelig velvære og høyere sikkerhet. Imidlertid mener Siau og Wang at KI-basert teknologi utgjør betydelig risiko for utviklere, brukere, samfunn og for menneskeheten, ettersom KI har lav forklaringsevne, fare for bias, problemer knyttet til datasikkerhet og personvern og andre etiske bekymringer (2020, 74). Etter hvert som KI tar en større plass i samfunnet, har det blitt et kritisk problem å finne ut hvordan man skal håndtere disse moralske og etiske bekymringene som Siau og Wang identifiserer. KI-etikk er et felt som har vokst fram for å svare på disse bekymringene (Kazim og Koshiyama 2021; Sieu og Wang 2020). Ifølge Kazim og Koshiyama (2021, 1) er KI-etikk et nytt felt og en undergruppe av digital etikk, som adresserer bekymringer reist av distribusjon og utvikling av nye digitale teknologier.

KI-etisk litteratur er et godt verktøy for å analysere norsk KI-etisk tilnærming, og se teknologien gjennom en kritisk linse. En av de mest utbredte tilnærmingene for å sikre etisk KI har vært å benytte retningslinjer som ofte bygges på etiske prinsipper (Kazim og Koshiuama 2018; Sieu og Wang 2020; Hickok 2020; Jobin, Ienca og Vayane 2018). Etter hvert som bruken av KI har blitt mer utbredt, og bekymringene har vokst, har et svar vært å

regulere teknologien. De siste årene har en rekke private og offentlige bedrifter og etater publisert KI-etiske prinsipper og retningslinjer (Hickok 2020). Denne prinsippbaserte tilnærmingen har som mål å gi veiledning og struktur for bruken av teknologi (Kazim og Koshiyama 2021, 6). Haalama og Kalliokoski (2022, 2) skriver i sin artikkel «AI ethics as applied ethics», at fra et filosofisk perspektiv representerer ulike formuleringer av KI-etikk, en forståelse av at KI-etiske retningslinjer er en form for anvendt etikk. KI-etiske retningslinjer evaluerer hvordan man kan konseptualisere, analysere og vurdere etisk relevante aspekter ved design samtidig som de legger fram metoder for styring og regulering av teknologi og utvikling (2). Forfatterne skriver at i KI-etiske rammeverk er det variasjoner i balanse mellom økonomisk lønnsomhet og sosial bærekraft, men at alle deler to viktige forutsetninger: felles vilje om å utvikle KI på en ansvarlig og moralsk måte, og en forståelse av hvordan retningslinjer er ment å forbedre verden (Haalama og Kalliokoski 2022, 2).

Kazim og Koshiyama (2021), Jobin, Ienca og Vayana (2019) og Hickok (2020) identifiserer flere sentrale KI-etiske prinsipper innenfor KI-etikken: *menneskets beste*, *sikkerhet*, *personvern*, *transparens*, *rettferdighet* og *ansvarlighet*. Dette er kun et utvalg av etiske prinsipper, men de som er mest gjentakende i forskning og i utgitte retningslinjer. Nedenfor vil jeg beskrive de ulike KI-etiske prinsippene hentet fra Kazim og Koshiyama (2021), sine beskrivelser.

Menneskets beste bygger i følge Kazim og Koshiyama (2021, 8) på etiske prinsipper om respekt for menneskeverd og inkluderer miljømessig, sosialt og psykisk velvære. En sentral problemstilling innen dette prinsippet er hvordan KI påvirker menneskers handlefrihet inkludert deres mentale autonomi og påvirkning på individet. Kazim og Koshiyama skriver at dette omfatter måter KI-systemer kan redusere menneskers rasjonelle kapasitet, enten direkte eller indirekte. I tillegg kan KI komme i konflikt med menneskelig handlefrihet, ifølge Kazim og Koshiyama, ettersom respekt for individet krever at samtykke er meningsfylt og informert – noe KI ofte ikke oppnår. Menneskets beste fra et samfunnsperspektiv omfatter identitet, tilhørighet og fellesskap, og inkluderer også de juridiske, politiske, demokratiske og økonomiske konsekvensene av KI. Dette stiller spørsmål ved bærekraft, partiskhet og rettferdighet (Kazim og Koshiyama 2021, 8).

Sikkerhet er knyttet til det etiske prinsippet om å forebygge skade, der skade forstås som negative effekter på menneskers velvære, inkludert de sosiale, miljømessige og psykologiske dimensjonene (Kazim og Koshiyama 2021, 8). Tilnærmingen innebærer først å identifisere risikoer og deretter redusere dem, med forebygging som et kjerneelement (8).

Personvern er i følge Kazim og Koshiyama (2021, 9) nært knyttet til offentlige og politiske krav om å beskytte individers personlige informasjon, basert på et tydelig skille mellom den private og personlige sfæren og den offentlige, politiske eller kommunale sfæren. Forfatterne understreker at den private og personlige sfæren krever et høyere nivå av respekt for individets rett til privatliv. Informert samtykke utgjør en sentral komponent i denne konteksten, fordi det sikrer at individer er klar over hvordan deres data samles inn, lagres og brukes. Samtidig foregår det en pågående debatt om verdien av personopplysninger og hvordan de økonomiske fordelene som genereres, bør fordeles rettferdig (Kazim og Koshiyama 2021, 9).

Ifølge Kazim og Koshiyama (2021, 9) er *transparens* forankret i prinsippet om åpenhet, som spiller en avgjørende rolle for å fremme tillit og ansvarlighet. Forfatterne beskriver åpenhet som både knyttet til beslutninger om bruken av KI-systemet og til hvordan systemet selv tar beslutninger. Førstnevnte handler om styring, mens sistnevnte omhandler en forklaring på hvordan automatiserte beslutningssystemer fungerer (Kazim og Koshiyama 2021, 9).

Rettferdighet er basert på det etiske prinsippet om menneskelig likhet, og er et tema som ofte diskuteres, der flere er uenig i begrepets betydning, hvilke definisjoner vi skal forplikte oss til (Kazim og Koshiyama 2021, 9). Dette mener Kazim og Koshiyama (2021, 9) er fordi det er flere ulike teorier – slik som korrigerende, distribuerende, prosedyremessige, substansielle og komparative teorier. Spørsmålet gjelder også hvilke rammer for rettferdighet som diskuteres, ifølge Kazim og Koshiyama, enten det dreier seg om rettferdighet i politiske fellesskap (som statsborgerrettigheter) eller universelle menneskelige bekymringer, og hvordan demografiske faktorer som kjønn, nasjonalitet, rase, natur, sosioøkonomisk bakgrunn og lignende defineres (Kazim og Koshiyama 2021, 9).

Ansvarlighet i etisk KI mener Kazim og Koshiyama (2021, 9) kan knyttes til anvendt etikk. Ansvarlighet omfatter ifølge forfatterne beslutningsprosesser, beslutningslogikk, fordeling av ansvar, utvikling og hvordan risikoer og skader skal måles og vurderes. Når man jobber ut ifra ansvarlighetsprinsippet mener Kazim og Koshiyama (2021, 9), at det er viktig å vite hvem som tar beslutninger, hvordan de blir tatt, og hvilke verktøy og systemer som benyttes for å måle dem. Forfatterne skriver også at fordeling av juridisk ansvar er sentralt innenfor prinsippet, der en viktig forutsetning for å fordele ansvar er modellenes forklarbarhet for å vite hvorfor systemene fatter sine beslutninger står sentralt for å vite hvor man skal plassere ansvaret (Kazim og Koshiyama 2021, 9).

2.3 KI-etikk og Etikk

Etikk blir forsket på innen mange ulike fagfelt. Siau og Wang (2020, 75) skriver at sammen med en økende interaksjon mellom mennesker, mennesker og dyr, mennesker og maskiner og til og med mellom maskiner har etiske teorier blitt anvendt innenfor en rekke områder som forretningsetikk, dyreetikk, militæretikk, bioetikk og maskinetikk. Forfatterne mener at studiet av etikk og etiske prinsipper er i kontinuerlig utvikling og tilpasser seg stadig nye utfordringer og kontekster, som for eksempel KI-etikk (Siau og Wang, 2020, 75), som denne oppgaven utforsker. Forfatterne beskriver etikk som et komplekst og sammensatt begrep. De definerer det som: "de moralske prinsippene som styrer atferden eller handlingene til et individ eller en gruppe individer" (Siau og Wang 2020, 75, egen oversettelse). Ifølge forfatterne er etikk et system av regler, prinsipper eller retningslinjer som skiller rett fra galt, der etikk i brede trekk kan beskrives som en disiplin som vurderer rett fra galt, samt moralske plikter og forpliktelser (Siau og Wang 2020, 75). Tasioulas (2022, 232) skriver at etikk er knyttet til eksistensen av menneskelige valg som vurderes gjennom et sett med verdier.

Berry (2020, 450) påpeker at mye av litteraturen om KI og etikk ofte retter seg mot enten konsekvensetikk eller deontologisk etikk. Ifølge Berry vektlegger deontologisk etikk at valg må tas i samsvar med moralske normer eller plikter, uavhengig av konsekvensene av handlingene. Dette perspektivet kan relateres til etterlevelse av KI-etiske prinsipper og retningslinjer, fordi disse er et sett med regler som gjør det mulig for andre å stille de som

bryter prinsippene til ansvar. Ifølge Berry (2020, 450) vektlegger dydsetikk den moralske karakteren til personen som handler, hvor handlinger vurderes ut fra dyder og laster. Dyd forstås som utviklingen av en god moralsk karakter. Gjennom denne tilnærmingen er en person dydig dersom vedkommende besitter praktisk visdom, noe som gjør dem i stand til å identifisere de moralsk relevante aspektene i en situasjon og velge riktig handling (450). Tasioulas (2022, 232) skriver at etikk er uunngåelig innenfor beslutningstaking om KI, fordi valg om hvordan man skal utvikle og distribuere teknologien kun kan forstås i lys av vårt forsøk på å følge etiske verdier. Alle former for regulering, i følge Tasioulas (2022, 233) impliserer valg som reflekterer etiske verdier og prioriteringer. Tasioulas mener at innhold i etiske standarder av mange blir tolket utelukkende som et spørsmål om rettferdighet, som kan knyttes til hvordan mennesker blir ulikt behandlet beskrevet som «algoritmisk urettferdighet» (2022, 233).

Tasioulas (2022, 233-234) skriver at etikk ofte blir oppfattet for å ha et for «snevert og individualistisk fokus», der man veileder individers personlige oppførsel framfor å påvirke de større institusjonelle og sosiale rammene innenfor hvor individers beslutninger tas. Denne oppfatningen utfordrer Tasioulas, som skriver at de fleste etiske verdier har konsekvenser for institusjoner og for sosial organisering. Tasioulas (234) mener at det er en vanlig oppfatning at etikk handler om normer som ikke kan håndheves gjennom lover eller rettssystemer, men som heller støttes av individuell samvittighet og samfunnets uformelle meninger. Men å begrense etikk til "myke" reguleringsformer er vilkårlig, skriver Tasioulas (234). Spørsmål om vi bør innføre lover og regulering, og hvordan disse skal håndheves, er ifølge Tasioulas et etisk spørsmål som involverer verdier om personlig frihet og rettferdighet. Rettferdighet har en lang tradisjon, på tvers av ulike ideologier, som noe som bør sikres gjennom lover og regler (235). Tasioulas argumenterer derfor for en bredere tilnærming til etikk i sammenheng med KI. Denne tilnærmingen bør omfatte alle former for regulering, fra individers personlige ansvar til juridiske rammeverk, og kan få betydelig innvirkning på hvordan maktstrukturer i samfunnet organiseres (236).

2.4 Kunstig intelligens

Dette delkapittelet fokuserer på de teknologiske aspektene som jeg anser som essensielle for å kunne besvare problemstillingen innenfor et komplekst og tverrfaglig felt. Dette innebærer en avgrensning av de tekniske detaljene til de som er mest relevante for den overordnede etiske diskusjonen. Det er viktig å ha en grunnleggende forståelse av teknologiske prosesser for å kunne foreta velinformerte etiske vurderinger, som er et av argumentene i denne oppgaven for å oppnå etisk bruk av KI.

2.4.1 Definisjoner

Kunstig intelligens er et bredt og omfattende begrep som spenner over en rekke disipliner, (Kommunal- og moderniseringsdepartementet, 9) Regjeringen skriver at på grunn av begrepets tverrfaglighet varierer definisjonene av KI mellom ulike fagfelt, og disse definisjonene kan også utvikle og endre seg i takt med teknologiske fremskritt (Kommunal- og moderniseringsdepartementet, 9). Ettersom denne oppgaven ser på KI i en norsk offentlig kontekst, har jeg valgt å inkludere regjeringens definisjon på KI hentet fra den nasjonale strategien for KI:

> «Kunstig intelligente systemer utfører handlinger, fysisk eller digitalt, basert på tolkning og behandling av strukturerte eller ustrukturerte data, i den hensikt å oppnå et gitt mål. Enkelte KI-systemer kan også tilpasse seg gjennom å analysere og ta hensyn til hvordan tidligere handlinger har påvirket omgivelsene» (Kommunal- og moderniseringsdepartementet 2020, 9).

For å illustrere hvor varierte definisjonene av KI kan være, har jeg inkludert en definisjon som samler både tekniske og filosofiske perspektiver hentet fra Lanestedt, Goodwin og Andersen (2023):

> «KI representerer ambisjonen om å etterligne de kognitive prosessene vi vanligvis assosierer med menneskelig tankegang – altså evnen til å resonnere, lære av tidligere erfaringer og fremfor alt ta informerte beslutninger basert på tilgjengelig informasjon. KI er med andre ord en kompleks algoritme, eller oppskrift, på hvordan data skal

> forstås. KI er også et resultat av historiske ønsker om å overføre noen av menneskets mest dyrebare evner til maskiner» (Lanestedt, Goodwin og Andersen 2023, 8).

De to definisjonene skiller seg ved at den første er teknisk og fokuserer på KI-systemers funksjon, som dataanalyse, tilpasning og måloppnåelse, mens den andre har en filosofisk tilnærming og vektlegger KI som en etterligning av menneskelig kognisjon. I dag finnes det ingen global konsensus eller allment akseptert definisjon av kunstig intelligens. Wang (2019, 1) hevder at det er urealistisk å forvente en felles definisjon, gitt intelligensens kompleksitet.

Ifølge Wang (2019, 1) er utfordringen med å bli enig om en felles definisjon knyttet til selve begrepet *intelligens*. Fra et filosofisk perspektiv er intelligens et vagt begrep i følge Coeckelbergh (2020) som refererer til Jensen mfl. (2018) sin definisjon av KI: «vitenskapen og konstruksjonen av maskiner med evner som anses som intelligente i henhold til standarden for menneskelig intelligens» (Jansen mfl., 2018, som referert i Coeckelbergh, 2020, 64). Dette eksempelet viser hvordan KI kan defineres som maskiner som handler som mennesker. Coeckelbergh påpeker at mange forskere innen KI foretrekker en mer nøytral definisjon som ikke nødvendigvis er knyttet til menneskelig intelligens, men heller relaterer seg til mål for generell eller sterk KI. Ulike oppfatninger og definisjoner av begrepet KI springer ofte ut fra varierende oppfatninger om hva det vil si å være intelligent. Selv om denne oppgaven ikke følger et filosofisk perspektiv og ser problemstillingen gjennom tilnærminger som transhumanisme og posthumanisme, ønsker jeg å få fram at det er flere perspektiver, og ikke kun den tilnærmingen jeg bruker i denne oppgaven.

2.4.2 Algoritmer, maskinlæring og dyplæring

KI-systemer tolker data fra sensorer, mikrofoner og kameraer, samt andre informasjonskilder. De analyserer dataene, tar beslutninger og utfører handlinger (Kommunal- og moderniseringsdepartementet 2020, 10). Ofte er KI en del av større IT-systemer. Vanlige bruksområder inkluderer datasyn («computer vision»), som kan brukes til blant annet ansiktsgjenkjenning og bildediagnostikk, og språkbehandling («Natural Language Processing»), som hjelper med å sortere dokumenter og trekke ut viktig informasjon, som sett i verktøy som ChatGPT. Robotikk er et annet viktig KI-område, brukt i utvikling av autonome kjøretøy som selvkjørende biler. KI er spesielt nyttig i å identifisere mønstre og avdekke

avvik, for eksempel i datasikkerhet og svindel (Kommunal- og moderniseringsdepartementet 2020, 11). For å kunne evaluere hvorvidt et KI-system er etisk forsvarlig, er det avgjørende å ha en grunnleggende forståelse av teknologien og dens mange bruksområder.

I alle applikasjonene nevnt over er algoritmer en del av de avanserte systemene. Cormen, Leiserson, Rivest og Stein (2022) beskriver algoritmer som «en hvilken som helst veldefinert beregningsprosedyre som tar en verdi, eller et sett med verdier som input, og produserer en verdi, eller et sett med verdier som output. En algoritme er dermed en sekvens av beregningstrinn som omdanner input til output» (2022, 5, egen oversettelse). Det finnes altså mange ulike former for algoritmer innenfor KI. En av de mest brukte formene er maskinlæring. Mahesh (2018) beskriver maskinlæring, som «den vitenskapelige studien av algoritmer og statistiske modeller som datasystemer bruker for å utføre en spesifikk oppgave uten å være eksplisitt programmert» (Mahesh 2018, 381, egen oversettelse).

Læringsalgoritmer, som anvendes innen maskinlæring, er integrert i mange av de applikasjonene vi benytter daglig. For eksempel i et Google-søk, har algoritmen lært å rangere nettsider for å gi brukeren de mest relevante resultatene. Mahesh (2018) skriver at maskinlæringsalgoritmer brukes til en rekke formål, inkludert datautvinning, prediktiv analyse og bildebehandling. En av fordelene med disse algoritmene er at når de har lært hvordan data skal behandles, kan de implementeres og brukes umiddelbart (381). Maskinlæring er derfor spesielt nyttig når man håndterer store datasett, da teknikken gjør det enklere å tolke og hente ut informasjon fra dataene. Den økte tilgjengeligheten av store datasett har ført til en økende etterspørsel etter maskinlæring (381). Det finnes mange forskjellige maskinlæringsalgoritmer, men to grunnleggende tilnærminger er overvåket læring (som regjeringen har vist bekymring til) og uovervåket læring (Kommunal- og moderniseringsdepartementet 2020, 11). Den primære forskjellen mellom disse tilnærmingene er at overvåket læring benytter merkede data for å forutsi utfall, mens uovervåket læring ikke gjør det (Delua 2021; Mahesh 2018).

Dyplæring («deep learning») utgjør en underkategori av maskinlæring og spiller i dag en sentral rolle i en rekke løsninger, inkludert bildebehandling, datasyn, talegjenkjenning og språkbehandling. Andre anvendelsesområder inkluderer utvikling av legemidler, anbefalingssystemer for musikk og film, behandling av medisinske bilder, personifisert

medisin og avviksdeteksjon innen ulike felt (Kommunal- og moderniseringsdepartementet 2020, 12). Enkelte dyplæringsalgoritmer kan beskrives som en «sort boks», ettersom det ofte er vanskelig å forstå de underliggende prosessene som forklarer hvordan en spesifikk inndataverdi fører til et bestemt resultat (Kommunal- og moderniseringsdepartementet 2020, 12). For å trene en algoritme trenger man data Whang, Roh, Song og Lee (2023) mener at prosessen krever nøye dataforberedelse, som innebærer innsamling, rensing og klargjøring av data slik at de kan brukes i treningen. Dette er ofte en kostbar og tidkrevende oppgave, fordi dataene må være av høy kvalitet og i riktig format for at modellen skal kunne lære effektivt. Rensingen kan inkludere å fjerne feil, mangler eller irrelevante elementer i datasettet. God dataforberedelse er avgjørende for å oppnå nøyaktige og pålitelige resultater fra maskinlæringsalgoritmene (791). I delkapittel 4.2.2, vil jeg redegjøre for dataens viktige rolle, og hvordan mangel på inkludering av mangfold i datagrunnlag kan videreføres og forsterkes gjennom automatiseringen.

2.5 Kunstig intelligens og humanistisk forskning

Humanistisk tenkning og forskning er viktig i debatter knyttet til KI. Chun og Elkins (2023, 149) argumenterer for at vi nå, mer enn noen gang, trenger humanister. Forfatterne påpeker at vi står ved et veiskille hvor «vi kan enten programmere våre menneskelige verdier inn i teknologien vi utvikler, eller risikere å miste vår humanitet til kunstig intelligens» (149, egen oversettelse). Teknologien har allerede betydelig innflytelse på mange aspekter av livene våre ettersom KI-modeller benyttes til å avgjøre hvem som får jobb, lån og tilgang til viktige ressurser. Dette er noen av årsakene til at Chun og Elkins (2023, 149) understreker viktigheten av å utvikle KI som gagner hele menneskeheten, i stedet for å prioritere smale økonomiske gevinster eller å minimere en abstrakt tapsfunksjon. Dette prinsippet er blitt formalisert som et eget felt innen KI, kalt menneskesentrert KI (HAI) (2023, 149).

Chun og Elkins (2023, 153) skriver at store KI-modeller er kontrollert av private selskaper som prioriterer aksjonærene, som ofte er i konflikt med menneskets beste. Et sentralt problem er at man har begrenset innsikt i hvordan modellene fungerer i praksis. Forfatterne mener at når KI-teknologi modnes og i økende grad påvirker alle aspekter av våre offentlige og private liv, haster det å få inn digitale humanister til å engasjere seg konstruktivt i design, utforming,

implementering og overvåking av KI-systemer. Som de nevner, bør vi «alle ha en rolle i å bestemme hvordan disse KI-modellene trenes og med hvilke data, for å vurdere hvordan de presterer og i hvilke populasjoner» (Chun og Elkins, 153, egen oversettelse).

KI-utviklere er ofte adskilt fra forskere og studenter som tar for seg viktige samfunnsmessige spørsmål. Derfor mener Frank mfl. (2019, 80) at det er ønskelig med økt forskningsinteresse mellom ulike fagområder og KI. Chun og Elkins (2023, 149) argumenterer for det samme, nemlig at KI-humanister, som har fått teknologisk innsikt gjennom KI-orienterte kurs i digital humaniora, kan tilføre verdifulle perspektiver til utviklingen av menneskesentrert KI. Forfatterne ser en trend der KI integreres i hele universitetsstrukturen, noe som bidrar til å viske ut grensen mellom teknologi og humaniora.

For å utvikle KI-systemer som gagner hele menneskeheten, og utfyller framfor å erstatte mennesket, har flere topprangerte informatikkskoler hentet inn eksperter fra humaniora for å takle problemstillingen (Dimock 2020, 450). En av disse er Stanfords Institute for Human-Centered Artificial Intelligence, som forsker på menneskerettet KI. Ved å følge en tverrfaglig tilnærming til problemstillinger rundt KI, tvinges viktige spørsmål fram – som hva det vil si å være menneske, og hvordan teknologi skiller seg fra menneskeheten. Ifølge John Etchemendy, en meddirektør for HAI-instituttet ved Stanford, vil KI «forvandle alle disipliner, spesielt humaniora – fordi det vil generere grunnleggende spørsmål som humaniora er best rustet til å svare på» (Etchemendy, referert i Dimock 2020, 450, egen oversettelse).

Dimock (2020) skriver at helt siden Aristoteles sin tid, har rasjonalitet vært et kjennetegn ved mennesker som skiller oss fra andre dyr. «Hva skjer når denne definerende egenskapen deles, eller til og med overskrides, av artefakter som vi mennesker lager?» (Dimock 2020, 451, egen oversettelse). Dette er spørsmål som menneskerettet KI ønsker å svare på med sin tverrfaglige tilnærming. Direktørene bak HAI-instituttet ved Stanford mener at om KI skal bidra til å tjene menneskehetens kollektive behov, så må menneskene som står bak systemene representere menneskeheten, noe som krever mangfold av tanke, på tvers av etnisitet, kjønn, kultur, alder, nasjonalitet og disipliner (Li og Etchemendy u.å., referert i Dimock 2020, 451). På samme måte mener Chun og Elkins (2023, 153) at viktige etiske spørsmål bør vurderes av et mangfold av perspektiver, ikke bare ingeniører. Forfatterne identifiserer et problem: man

trenger tverrfaglig digital kompetanse for å kunne bidra konstruktivt til de store etiske spørsmålene, noe som vil kreve nye former for utdanning.

Kapittel 3: Metode

Det har de siste årene blitt tydelig at vi må ta etiske hensyn når vi jobber med teknologi, noe som er en anerkjennelse av at teknologien kan gi negative konsekvenser for enkeltmennesket og samfunnet. Selv om mange er enig i at teknologien må brukes på en etisk, forsvarlig og god måte, er det ikke enighet i hvordan dette fungerer i praksis eller hva «etisk» bruk er. For å svare på oppgavens problemstilling om hvordan vi kan sikre at etiske hensyn står sentral i implementeringen av kunstig intelligens i norsk offentlig sektor, vil jeg i dette kapittelet gjøre rede for de metodiske valgene jeg har gjort, og begrunne disse valgene opp mot problemstillingen. Kapittelet beskriver egen førforståelse, valg av en mikset kvalitativ metode bestående av en litteraturundersøkelse og intervju, datainnsamling, transkribering, tematisk analyse, etiske hensyn, begrensninger og til slutt oppgavens validitet og reliabilitet.

3.1 Kvalitativ metode

For å svare på oppgavens problemstilling, har jeg valgt å følge en kvalitativ metode som består av en litteraturundersøkelse kombinert med semi-strukturerte intervju. Fordelen med kvalitative metoder er at de gjør det mulig å fange opp meninger og opplevelser som ikke lar seg tallfeste eller måle (Dalland 2022, 54). Ved å benytte kvalitativ metode har jeg muligheten til å gå i dybden i oppgavens problemstilling og å utforske nye funn gjennom datainnsamlingen. Dalland skriver at de som benytter seg av kvalitativ metode kalles «tolkere» i motsetning til «tellere» som kjennetegner kvantitativ metode (Dalland 2022, 54). Begrunnelsen bak valget å kombinere intervju og litteraturundersøkelse er at det finnes lite forskning som tar for seg KI-etiske bekymringer i en nasjonal kontekst, framfor mye som tar for seg den globale. På et internasjonalt nivå er det mye grunnforskning som ser på etiske bekymringer til teknologien, hvor flere aspekter kan knyttes til norsk implementering. Likevel er det er et betydelig gap i å konkretisere KI-etiske bekymringer opp mot en norsk kontekst, som er hvorfor jeg valgte å også ha intervjuer.

Før jeg går videre inn i metodekapittelet ønsker jeg å redegjøre for min egen posisjon innenfor oppgavens tematikk, noe Dalland (2022, 60-61) skriver er viktig i kvalitativ forskning fordi alle har fordommer eller førforståelser med seg inn når man starter en undersøkelse. I tidligere fag og oppgaver på universitetet har urettferdig bruk av KI vært en interesse. Gjennom et

humanistisk perspektiv har jeg lært å sette spørsmålstegn til teknologiens påvirkning på samfunnet og enkeltmennesker. I denne masteroppgaven ønsker jeg å ha et så objektivt syn på både problemstillingen og datainnsamlingsprosessen som mulig, men jeg anerkjenner likevel at det er vanskelig å unngå at subjektive tanker rundt tematikken reflekteres i oppgaven. Overordnet har oppgaven et humanistisk perspektiv som man vil se gjennom oppgaven. Likevel krever oppgavens tverrfaglige problemstilling at jeg åpner opp for andre perspektiver og ideer.

3.1.1 Litteraturundersøkelse: Utvalg og gjennomføring

I begynnelsen av mastergradsprosjektet ønsket jeg å basere litteraturundersøkelsen utelukkende på fagfellevurderte forskningsartikler. Etter hvert som jeg gjorde research til oppgaven, merket jeg at det var mye grunnforskning som så på problemstillingen fra en internasjonal kontekst, men lite forsking som tok for seg den norske konteksten. Siden datainnsamlingen skal følge en systematisk prosess, valgte jeg å strukturere litteraturundersøkelsen i to deler: én som belyser temaet i en nasjonal sammenheng, og én som ser det fra et internasjonalt perspektiv. De to delene lar meg sette forskjellige krav til litteraturen. Den internasjonale delen ser problemstillingen fra et rent forskingsperspektiv, mens den nasjonale delen ser også i stor grad på forskning, men inkluderer også styringsdokument publisert av det offentlige, samt kilder hentet fra kronikker og artikler som er skrevet av forskere, men som ikke nødvendigvis er fagfellevurderte.

Litteraturen brukt i denne oppgaven er funnet ved hjelp av akademiske søkemotorer som Research Gate, Google Scholar, Universitetet i Bergen sin digitale bibliotek-søkemotor Oria, Idunn, samt lærebøker ved studieprogrammet Digital kultur. For å finne relevant litteratur benyttet jeg meg av søkeordene: «AI-ethics», «AI-ethics guidelines», «AI-ethics theory», «AI-ethics benefits» og «AI-ethics risks». De samme ordene er brukt med norsk oversettelse for å finne norsk litteratur. I den første fasen av datainnsamlingen leste jeg innledningen eller abstraktene til aktuelle artikler for å evaluere om de var verdifulle for min problemstilling. Under denne prosessen undersøkte jeg om den internasjonale litteraturen var fagfellevurdert forskning, og om den nasjonale litteraturen enten var basert på forskning eller utgitt av offentlige aktører, som departementer og direktorater. Litteratur som både var relevant for problemstillingen og oppfylte kriteriene ble lagret i en egen mappe på datamaskinen. Deretter

leste jeg gjennom artiklene og skrev korte sammendrag av hovedtemaene og argumentene som ble presentert for å gjøre datamaterialet klart for tematisk analyse, som beskrives i delkapittel 3.2.

3.1.2 Intervju: Utvalg og rekruttering

Under utvelgelsen av intervjuobjekt valgte jeg å følge det Dalland kaller for et strategisk utvalg (Dalland 2022, 59). Ettersom det var lite litteratur tilgjengelig som tar for seg den norske konteksten opp mot en etisk god implementering av teknologien, valgte jeg å intervjue eksperter innen feltet for å bidra til å tette dette gapet. Ekspertene jeg valgte å intervjue har bakgrunn fra akademia - et valg jeg tok for å samle informasjon fra et høyt faglig nivå. Begrepet «ekspert» er i sin natur subjektivt, i denne oppgaven har jeg operasjonalisert det som en person med faglig kompetanse og akademisk ansettelse. Kriteriene for intervjuobjektene er at de har doktorgrad, fast ansettelse og forsker på tema som er relevant for oppgavens problemstilling. Et valg jeg tok i denne prosessen var å inkludere intervjuobjekt med både humanistisk og teknisk bakgrunn. Selv om dette er en oppgave som tar utgangspunkt i humanistiske perspektiver, er oppgavens problemstilling tverrfaglig, og derfor ville jeg inkludere tekniske perspektiver for å få et mer nyansert svar.

I rekrutteringsprosessen gikk jeg inn på nettsider som viser ansatte hos ulike utdanningsinstitusjoner innen høyere utdanning. Ut fra mine satte kriterier for intervjuobjekt, fant jeg kandidater som fylte utvalgskriteriene til tittel, forskning og bakgrunn. Når jeg hadde en liste med aktuelle kandidater var første steg å kontakte dem via e-post. I e-posten til intervjuobjektet fortalte jeg om prosjektet og spurte om dette var noe kandidaten ønsket å være med på. I e-posten la jeg ved et samtykkeskjema (se vedlegg 2) som gikk dypere inn på oppgavens formål og beskrev hva det innebærer å delta, hvordan data og personvern ville bli behandlet og ivaretatt, samt deltakerens rettigheter og en samtykkeerklæring. Når kandidater ønsket å være med, avtalte vi dato og klokkeslett for intervjuet over e-post.

3.1.3 Intervju: Gjennomføring

I forkant av intervjuene utarbeidet jeg en intervjuguide (se vedlegg 1) bestående av ti spørsmål. Spørsmålene ble rangert og sortert i rekkefølge etter viktighetsgrad, for å sikre at jeg fikk stilt de viktigste spørsmålene til å hjelpe meg å svare på min problemstilling. Ved å velge en semi-strukturert metode for gjennomførelsen av intervjuene ga dette meg rom til å gå inn i andre tema og problemstillinger som kom fram under intervjuene (Kvale og Brinkmann, 2009, 47). I forkant av intervjuet valgte jeg å ikke lese intervjuobjektenes forskningsartikler og bøker, for å redusere en mulig påtvungen forutinntatt forestilling på mine forventninger til intervjuet. Intervjuene ble holdt over Zoom og hadde varighet på rundt én time. Under forberedelsene til intervjuet leste jeg gjennom egne notater og spørsmål samt deltakernes samtykkeskjema. Under disse forberedelsene innså jeg at samtykkeskjemaene ikke eksplisitt presiserte at det ville bli tatt lydopptak av samtalen. Derfor var det viktig for meg å igjen informere og være tydelig på at jeg ønsket å ta lydopptak av samtalen, og spørre om informantene samtykket til dette, før intervjuene startet. Begge informanter ga nytt samtykke, og har signert samtykkeskjema. Før jeg startet intervjuet, spurte jeg også intervjuobjektene om de hadde spørsmål til intervjuet og om deres deltakelse. Lydopptakene ble slettet med en gang de var ferdig transkribert.

3.1.4 Transkribering

Dataen innhentet gjennom intervjuene ble transkribert manuelt på PC. Første steg i denne prosessen var å høre gjennom lydopptakene for å få en oversikt over datamaterialet før jeg startet selve transkriberingen. På bakgrunn av observasjonene gjort ved å høre på lydopptakene lagde jeg kodeord for ord som gjentok seg ofte, slik som KI (kunstig intelligens) og KI-E (kunstig intelligens etikk). Under selve transkriberingsprosessen noterte jeg hvem som sa hva og markerte hvor jeg selv stilte spørsmål, for å gjøre det lettere å bearbeide transkripsjonen i etterkant. I transkriberingen ble deltakerne anonymiserte for å ivareta personvern. Jeg valgte å ikke inkluderte latter, pauser, endring i tonefall og andre former for verbal- og ikke verbal kommunikasjon, ettersom jeg ikke vurderer det som relevant for å svare på min problemstilling. I mange tilfeller er slik kommunikasjon viktig å dokumentere, blant annet når man skal dokumentere følelsesmessige signaler, men i denne sammenhengen var jeg kun ute etter deltakernes utsagn. For å bearbeide datagrunnlaget fra intervjuene leste

jeg nøye gjennom transkripsjonen flere ganger. Det neste steget var å bearbeide transkripsjonen om til en mer leselig form for tekst, hvor jeg la til beskrivelser som: «informanten forteller, beskriver og sier», for å få teksten mer helhetlig.

3.2 Tematisk analyse

Kvale og Brinkmann skriver at det å analysere betyr «å dele noe opp i biter og elementer» (2015, 219). Når jeg analyserte datamaterialet mitt fra intervju og litteraturundersøkelse, var det overveldende å formulere data til tekst. Med mange sider transkripsjon og et stort datagrunnlag med litteratur var det vanskelig å vite hvor jeg skulle starte. For å bearbeide det store datagrunnlaget valgte jeg å gjøre en tematisk analyse. Anker (2020, 28) beskriver tematisk innholdsanalyse som den vanligste analyseformen i masterprosjekter, blant annet fordi analyseformen er god når man skal orientere seg i et stort datamateriale. Når man bruker tematisk innholdsanalyse av kvalitativt materiale, mener Anker (2020, 28) at hvor mange ganger et begrep gjentar seg er mindre viktig enn konteksten et begrep brukes i. Dette perspektivet har jeg hatt i tankene gjennom hele analyseprosessen.

3.2.1 Tematisk analyse: Litteratur

For å gjøre datamaterialet fra litteraturundersøkelsen enklere å analysere, lagde jeg en liste med sammendrag over alle tekstene og artiklenes argument. Da jeg skulle tematisere litteraturen, startet jeg med å identifisere tema jeg på forhånd av analysearbeidet mente ville være sentralt, og limte sammendragene inn under temaet som passet best. Ved å ha en kort oppsummerende tekst ble det enklere for meg å prøve ulike oppbygninger av tema, noe som fikk meg til å se nye sammenhenger og utvikle nye tema. Da jeg så meg fornøyd med tema og tekstenes plassering i oppgaven, startet prosessen med å skrive oppgaven.

3.2.1 Tematisk analyse: Intervju

For den tematiske analysen av intervjuene, skriver Dalland (2022) at man danner et grunnlag for analysen allerede når man utvikler intervjuguiden. Før man intervjuer har man tanker om hva man vil vite mer om gjennom intervjuet, fordi svarene skal hjelpe med å belyse problemstillingen (Dalland 2022, 94). Grunnarbeidet med intervjuguiden (vedlegg 1) satte de

første føringene for å tematisere datagrunnlaget. Prosessen med å tematisere intervjuene var annerledes enn litteraturen, fordi intervjuene var to lange og sammenhengende tekster. Jeg startet analysearbeidet med å sette intervjuene opp mot hverandre i en kronologisk rekkefølge som fulgte intervjuguiden. Neste steg var å identifisere gjentakende tema, dette gjorde jeg fordi jeg ønsket at temaene skulle komme fra materialet framfor å presse materialet inn i forhåndsbestemte tema. Videre tok jeg utgangspunkt i de potensielle temaene og limte inn tekstutdrag fra intervjuene, for å se hvordan de passet inn. Å plassere tekstutdragene inn i tema var utfordrende, fordi noen utdrag passet inn mange steder, mens andre var vanskeligere å plassere. Fordi intervjuene fulgte en semi-strukturert metode ble det ulikheter i datamaterialet, som er positivt, men som gjorde analysearbeidet mer krevende. Da jeg skulle tematisere intervjuene ønsket jeg ikke å plassere intervjuobjektenes svar ut av kontekst. Så selv om et svar var mer relevant innenfor et annet tema, ønsket jeg å bevare den rette konteksten for å ikke gi svaret deres en annen betydning. Dette var hele veien med i tankene gjennom analysearbeidet. Dette førte til at et par av temaene ble mer generelle enn andre, men et valg jeg tok bevisst for å bevare stemmene til informantene.

3.3 Kvalitetsvurdering av metode

I dette underkapittelet har jeg gjort en kvalitetsvurdering av denne studiens metode. Jeg viser til hvilke etiske hensyn jeg har tatt, og vurderer begrensinger, validitet og reliabilitet ved min metode.

3.3.1 Etiske hensyn

Det empiriske grunnlaget bygger blant annet på data fra intervjuer, og det er derfor spesielt viktig å vurdere forskningsetikk og begrunne forskningsetiske valg for å ivareta persondata. Som jeg har beskrevet tidligere i metodekapittelet, sendte jeg ut et samtykkeskjema i forkant av intervjuene (se vedlegg 2). Samtykkeskjemaet er utarbeidet på bakgrunn av en mal publisert av Sikt (Sikt u.å.), som er en tjenesteleverandør for kunnskapssektoren for forskningsetikk og personvern. Jeg vurderte derfor Sikts mal som et godt grunnlag, fra en sikker kilde, til mitt eget samtykkeskjema. Samtykkeskjemaet mitt opplyser om oppgavens formål, hva det innebærer for intervjuobjektene å delta, og informasjon om at opplysninger og data kun vil bli brukt for formålene informert om i samtykkeskjemaet. Det står også at

deltakere kan trekke seg når som helst, og at alle intervjuobjektene i utgangspunktet er anonymiserte. Det siste punktet er fordi jeg ønsket å publisere navn, institusjon og tittel på informantene for å styrke deres troverdighet som ekspert-stemmer i møte med de akademiske tekstene i litteraturundersøkelsen. Én av informantene samtykket til at navn, institusjon og tittel skulle publiseres, mens én ønsket å anonymiseres. For å ivareta mine informanters personvern og rettigheter, som er det viktigste i forskningsprosjektet, valgte jeg derfor å anonymisere alle informantene.

Prosjektet er godkjent i RETTE, som er Universitetet i Bergens «system for oversikt og kontroll med behandling av personopplysninger i forsknings- og studentprosjekter» (Universitetet i Bergen, 2024). Alle opplysninger og data samlet gjennom intervjuene har vært lagret på privat datamaskin med kodelås og slettes ved prosjektets slutt.

3.3.2 Begrensninger

Kvalitativ metode er en god metode når man ønsker å gå i dybden på oppgavens tematikk, og på samme tid ha fleksibilitet til å utforske gjennom datainnsamling (Dalland 2022, 54). Selv om jeg argumenterer for at valget av metode for denne oppgaven er passende, er det viktig å adressere begrensninger som kan oppstå på bakgrunn av denne tilnærmingen, og problemer som har dukket opp underveis i arbeidet med oppgaven.

Den første begrensningen jeg vil vise til, kommer på bakgrunn av de strenge utvalgskravene jeg satte for intervjuobjektene i denne oppgaven. Ved å sette krav til at intervjuobjektet skal ha doktorgrad innen forskning som er relevant for denne oppgavens problemstilling, og at de skal ha en humanistisk eller teknisk bakgrunn, ble resultatet at det ikke var mange kandidater som fylte alle kravene. I tillegg til dette er akademikere svært travle personer med trange timeplaner, så av alle de 19 kandidatene jeg tok kontakt med, var det kun to som ønsket og hadde mulighet til å stille til intervju. Framgangsmåten min, der jeg undersøkte ansattsider hos utdanningsinstitusjoner for å identifisere potensielle informanter, kan ha ført til at flere relevante kandidater ikke ble oppdaget. Konsekvensen av få informanter i det empiriske grunnlaget kan være en mindre nyansert analyse og drøfting, fordi oppgavens funn baseres på innsikt fra få personer framfor et større utvalg. Samtidig skriver Dalland (2022, 81) at det ikke er hensiktsmessig å intervjue for mange, og at få gode intervjuer kan gi mye innsikt. Dette har

vært sant for meg: selv om mine utvalgskriterier satte begrensninger, ga intervjuene meg likevel god innsikt i oppgavens tematikk fra et ekspert-nivå, og hjalp meg med å svare på oppgavens problemstilling. Et forskningsintervju som følger en kvalitativ metode, sikter mot å gå i dybden, og det er nettopp det mine intervjuer har gjort.

Den andre begrensningen jeg ønsker å vise til er at oppgaven overordnet ser på problemstillingen gjennom humanistiske perspektiv, selv om oppgavens problemstilling og tematikk er tverrfaglig, som jeg identifiserte innledningsvis i dette kapittelet. Når man bygger opp det empiriske grunnlaget, som i denne oppgaven er både litteraturundersøkelse og intervju, kan utvalget av data bli formet av linsen man ser gjennom. Selv om jeg har forsøkt å være en nøytral aktør gjennom intervjuprosessen, litteraturprosessen og analyseprosessen, så har linsen jeg ser gjennom bidratt til å forme oppgaven. Det er viktig at en masteroppgave som skrives innenfor humaniora beholder det humanistiske perspektivet, men likevel anser jeg det som viktig å få fram at tematikken er tverrfaglig, fordi andre viktige perspektiver kan ha fått mindre plass, som også kan bidra til en mindre nyansert analyse og drøftelse. Selv om jeg beskriver dette som en begrensning, er det også en styrke fordi humanistiske perspektiver ofte kan mangle i debatter rundt KI selv om det er sterkt behov for dem, som beskrevet i kapittel 2.

3.3.3 Validitet og reliabilitet

Validitet peker på at datagrunnlaget som måles må ha relevans og være gyldig innenfor tematikken oppgavens problemstilling (Dalland 2022, 43). Derfor er det viktig å gjøre rede for valg og vurderinger i denne prosessen. Dette er spesielt viktig i utviklingen av intervjuguiden, at spørsmålene man stiller bidrar til å belyse problemstillingen. I tillegg til å være relevant, må datagrunnlaget også være pålitelig (Dalland 2022, 63). Gjennom datainnsamlingsprosessen sørget jeg for å kun bruke akademiske kilder, og styringsdokumenter. Jeg valgte å intervjue akademiske eksperter innenfor problemstillingens felt, for å sikre innsikt fra et høyt faglig nivå. Når jeg utviklet intervjuguiden, tenkte jeg nøye gjennom hvordan spørsmålene mine kunne bidra til å svare på problemstillingen. Data samlet fra litteratur hadde strenge kriterier for å sikre relevante og gode kilder. Den semi-strukturerte intervjuguiden ga meg og intervjuobjektene rom for å snakke fritt rundt spørsmålene, men også til å utforske ny innsikt i feltet, som var målet med intervjuene, ved å stille oppfølgingsspørsmål.

Reliabilitet ser på oppgavens pålitelighet. Det handler om at målinger utføres korrekt, og at mulige feilmarginer fremlegges (Dalland 2022, 43). Pålitelighet er viktig for å sikre kvalitet innen forskning, og handler om hvorvidt man kan stole på arbeidet som blir presentert. Pålitelighet reflekteres gjennom å beskrive prosesser som valg av forskning, egen bakgrunn og førforståelse, gjøre rede for metode for datainnsamling og vise til feilkilder eller andre ting som kan ha påvirket resultatet (Dalland 2022, 58). Reliabilitet handler også om hvorvidt forskningsresultatene er konsekvente og troverdige, som vil si om et resultat kan reproduseres på et senere tidspunkt av andre forskere (Kvale og Brinkmann 2015, 276). Likevel er det vanskelig i kvalitativ forskning å sikre reliabilitet. Det empiriske grunnlaget følger både semi-strukturerte intervju, og en litteraturundersøkelse. Litteraturundersøkelser er god i pålitelighet i motsetning til intervju der det er lite sannsynlig at en annen forsker hadde fått akkurat det samme resultatet som min forskning. Den tematiske analysen av hele det empiriske grunnlaget har også lav reliabilitet, fordi det er lite sannsynlig at en annen forsker hadde tematisert dataene på helt lik måte som meg. For å styrke oppgavens pålitelighet har jeg derfor gjennom dette kapittelet forsøkt å beskrive mine valg, prosesser og begrensninger i detalj.

Kapittel 4: Litteraturundersøkelse

Målet med litteraturundersøkelsen er å belyse sentrale teorier, debatter og diskurser knyttet til etisk implementering av KI i norsk offentlig sektor. Presentasjonen av litteratur er basert på en tematisk analyse. Litteraturundersøkelsen er delt inn i to deler: én del som fokuserer på KI i en nasjonal kontekst (delkapittel 4.1) og én del som tar for seg internasjonale perspektiver hentet fra forskning (delkapittel 4.2).

4.1 Litteraturundersøkelse del 1: Nasjonalt nivå

I første del av litteraturundersøkelsen vil jeg se nærmere på implementeringen av KI i norsk offentlig sektor. I oppgavens teoretiske bakgrunn delkapittel 2.1 redegjorde jeg for regjeringens KI-etiske tilnærming. I de kommende delkapitlene har jeg undersøkt utbredelsen av KI-løsninger, hvordan de anvendes, og hvilken kritikk som har oppstått knyttet til deres implementering. I tillegg vil jeg gjenfortelle relevante lovverk og anbefalinger for hvordan KI kan innføres på en måte som er både etisk og hensiktsmessig i en norsk kontekst, gjennom forskning og offentlige dokumenter.

4.1.1 Status på implementering av KI i norsk offentlig sektor

I norsk statsforvaltning og offentlig sektor er bruk av kunstig intelligens i hurtig vekst. Ifølge Regjeringen (Kommunal- og moderniseringsdepartementet 2020) implementeres KI i Norge med et ønske om å skape effektive og produktive løsninger, som vil kunne føre til effektivisering av prosesser og arbeidsoppgaver og mer brukerrettede tjenester (5). Om man utnytter mulighetene teknologien bringer, kan det gi en rekke positive virkninger for norsk offentlig sektor og forvaltning ifølge Landestedt, Goodwin og Andersen (2023, 9). Forfatterne mener at det er ingen tvil om at den norske statsforvaltningen er på vei inn i det de kaller en «innovativ periode», hvor det stadig forskes på hvor og hvordan KI-teknologi kan tas i bruk (2023, 9).

De siste årene har det blitt mange offentlige aktører som ønsker å ta i bruk KI-teknologi. I 2023 publiserte NORA (Norwegian Artificial Intelligence Research Consortium) og Digitaliseringsdirektoratet (u.å.) en oversikt over aktuelle KI-prosjekter i offentlig sektor. Da oversikten ble publisert i 2023 inneholdt den allerede 135 prosjekter, hvor 40% stammer fra helsesektoren og 24% fra statsforvaltningen. Det er riktignok verdt å merke seg at denne

oversikten ikke er komplett, og at ikke alle pågående prosjekter er inkluderte. Flertallet av prosjektene er fortsatt i prøvestadier og har enda ikke nådd brukerstadiet (Digitaliseringsdirektoratet u.å).

KI har mange bruksområder. I Lanestedt, Goodwin og Andersen sin artikkel «Tid for en (mer) intelligent statsforvaltning?» (2023, 11-12) redegjør forfatterne for mange nyttige bruksområder for KI i offentlig sektor og forvaltning. Forfatterne mener blant annet at KI-teknologi kan automatisere og effektivisere offentlig saksbehandling og dokumentasjon, som kan bidra til å skape mer nøyaktig og effektiv saksbehandling og rapportskriving. Andre bruksområder foreslått av forfatterne er å bruke KI for å avdekke avvik, som beslutningsstøtte, eller for å utvikle KI-modeller som kan bidra til å styrke responsevnen – som å overvåke epidemier. Forfatterne mener også at chatbots har et stort potensial, fordi de kan gi brukere raske svar på spørsmål. Et siste bruksområde forfatterne viser til er ansiktsgjenkjenning og biometrisk identifikasjon som de skriver kan bidra til økt sikkerhet (Lanestedt, Goodwin og Andersen 2023, 11-12).

Å implementere KI både i det offentlige og private kan være vanskelig, fordi det er utfordrende å navigere seg gjennom hvordan man skal ta i bruk teknologien på en etisk, trygg og lovlig måte, som jeg ser nærmere på i kapittel 5 og 6. For å bistå bedrifter og offentlige tjenester til å ta i bruk KI på en god og sikker måte har Datatilsynet (2020) de siste årene utviklet og jobbet med et prosjekt de kaller den *regulatoriske sandkassen.* Datatilsynet (u.å.) skriver at sandkasseprosjektet er et kontrollert testmiljø for virksomheter som ønsker å eksperimentere med ny teknologi, så de kan utvikle nye tjenester og produkter med oppfølging fra myndighetene. Et mål for prosjektet er å styrke samarbeid mellom aktører og myndigheter. Ifølge Datatilsynet (u.å.) kan dette bidra til å gjøre det lettere å identifisere problemer og risikoer og bygge gode og trygge løsninger. Det kan også bidra til innovasjon med godt personvern, og utvikle etisk og ansvarlig KI (Datatilsynet u.å.). Datatilsynet ønsker at virksomheter skal få større forståelse knyttet til regulatoriske krav, og at myndigheter skal få økt forståelse for nye tekniske løsninger (Datatilsynet u.å.). Sandkassen tar utgangspunkt i tre hovedprinsipper som er hentet fra EU-kommisjonens retningslinjer for pålitelig kunstig intelligens: *lovlig, etisk og sikker* (Datatilsynet 2022, HLEG 2019).

Det første prosjektet som gikk gjennom Datatilsynets sandkasse, var NAVs prøveprosjekt for bruk av KI innen saksbehandling (Datatilsynet 2022). NAVs KI-verktøy skulle hjelpe saksbehandlere å forutsi hvor lenge mennesker vil være sykemeldt, for å videre kunne avgjøre hvem som ville ha behov for et oppfølgingsmøte to måneder fram i tid. Datatilsynets (2022) sin sluttrapport viste til flere utfordringer for personvern og rettferdighet, fordi forebygging av skjeve utfall krevde mer data. For å kunne avdekke og motvirke diskriminering krevde det nye metoder hvor man behandlet mer og nye former for personopplysninger som ikke allerede var en del av datagrunnlaget til KI-verktøyet (5), noe som utfordrer regjeringens prinsipp om dataminimering (Kommunal- og moderniseringsdepartementet, 57). En annen bekymring i Datatilsynets (2022) rapport var knyttet til innsikt og forståelse i KI-verktøyets virkemåte (15). Datatilsynet beskriver det som viktig at veiledere får god opplæring og instrukser i hva algoritmen skal brukes til og hvordan den fungerer. Dette kan ifølge Datatilsynet bidra veiledere i å vurdere algoritmenes utfall på et trygt og selvstendig grunnlag, og gjøre det mulig å avdekke feil, uønsket forskjellsbehandling og diskriminering (Datatilsynet 2022, 15). Økt digital kompetanse er sentralt for å sikre etisk bruk av KI-modeller, som jeg vil diskutere videre i oppgavens diskusjon.

4.1.2 Kritiske observasjoner av KI i norsk offentlig sektor

Selv om KI-teknologi kan tilby betydelige fordeler for offentlig forvaltning, fremhever Lanestedt, Goodwin og Andersen (2023, 12) behovet for «etiske, verdimessige og praktiske overveielser og vurderinger» for å håndtere utfordringene som teknologien bringer med seg. Ifølge forfatterne er den mest presserende utfordringen knyttet til hvordan KI påvirker arbeidsstyrken. Fremover vil flere statsansatte ha oppgaver relatert til algoritmer, data, utvikling og systemforvaltning, som forfatterne understreker vil kreve kompetanseheving blant ansatte med annen fagbakgrunn. Lanestedt, Goodwin og Andersen (2023, 12) peker også på at statsforvaltningen går inn i en periode preget av omfattende omorganisering, med endringer i arbeidsprosesser og måten det arbeides på. Derfor argumenterer forfatterne for at tettere samarbeid mellom virksomheter og på tvers av sektorer vil være avgjørende i tiden som kommer (12). I den nasjonale strategien for KI har regjeringen lagt vekt på behov for økt digital kompetanse og teknologiforståelse der regjeringen oppfordrer til tilrettelegging for å oppnå utvidet faglig kompetanse innen KI, og skape muligheter for at arbeidstakere kan ta videreutdanning (Kommunal- og moderniseringsdepartementet 2020, 43).

I forbindelse med implementeringen av KI-teknologi i norsk offentlig sektor identifiserer Lanestedt, Goodwin og Andersen (2023, 12–13) fem sentrale diskusjonstemaer som må adresseres: *transparens*, *bias* og *diskriminering*, *datasikkerhet* og *personvern*, *ansvarlighet* og *ansvarsplassering* og *lovregulering*. Lanestedt, Goodwin og Andersen (2023, 12) beskriver risikoen for redusert *transparens* som «en av de mest fremtredende bekymringene knyttet til KI.» Ifølge forfatterne kan det være svært utfordrende, og ofte umulig for mennesker å forstå hvordan og hvorfor en algoritme fatter beslutninger (12), som er hvorfor algoritmer ofte blir omtalt som en «svart boks» (Kommunal- og moderniseringsdepartementet 2020, 12) I norsk offentlig sektor er åpenhet et grunnleggende prinsipp og gjeldende lovverk krever at enkeltpersoner skal ha innsyn i hvordan beslutninger som påvirker dem, tas (Kommunal- og moderniseringsdepartementet 2020, 5). Lanestedt, Goodwin og Andersen (2023, 12) påpeker at dette prinsippet utfordres når maskiner overtar beslutningsprosesser. KI-verktøy kan være komplekse og vanskelige å tolke, noe som kan føre til redusert transparens rundt beslutningstaking. For å ivareta norske verdier som åpenhet og tillit, argumenterer forfatterne for at KI-systemer må utformes slik at de er så transparente og forståelige som mulig (12). Dette er viktig for å opprettholde tilliten i samfunnet og sikre at prinsippet om åpenhet forblir intakt (Kommunal- og moderniseringsdepartementet 2020, 5).

Lanestedt, Goodwin og Andersen (2023, 12) fremhever også risikoen for *bias* og *diskriminering* som en sentral bekymring som stammer fra hvordan data brukt for å trene KI-modeller ofte inneholder skjevheter og fordommer. For å unngå at algoritmer reproduserer eller forsterker eksisterende ulikheter, understreker forfatterne viktigheten av at data er representativt (12). Dette vil ifølge forfatterne, kreve kontinuerlig justering og overvåking av systemene, en oppgave for «eventuelle» fremtidige tilsynsmyndigheter (13). *Bias* og *diskriminering* vil jeg gå dypere inn på i del 2 av litteraturundersøkelsen, ettersom dette er et omdiskutert tema også på internasjonalt nivå

Datasikkerhet og *personvern*, er ifølge Lanestedt, Goodwin og Andersen (2023, 13), en annen sentral bekymring, ettersom den norske staten håndterer store mengder sensitive data om sine innbyggere. Å ivareta personvernet, vil innebære å ta hensyn til KI-systemers sårbarhet for angrep, fordi det kan få alvorlig konsekvenser for personvernet om sensitiv data lekkes (13). *Ansvarlighet* og *ansvarsplassering* er også en viktig diskusjon, ifølge forfatterne, fordi det kan

bli utfordrende å delegere ansvar når noe går galt. Er det KI-algoritmen, treningsdataene eller systemutviklere som skal stå ansvarlig, og hvordan ser dette ut i praksis, stiller forfatterne spørsmål til (13). Forfatterne skriver at det er avgjørende med en tydelig avklaring for spørsmål om ansvarlighet for å opprettholde rettssikkerheten i det norske samfunn, og beskriver denne problemstillingen som en «av de vanskeligste temaene å komme til bunns i» (13). Den siste bekymringen Lanestedt, Goodwin og Andersen (2023, 13) redegjør for er knyttet til lovregulering og EU sin KI-forordning, som jeg vil diskutere nærmere i 4.1.3.

Broomfield og Lintvedt sin artikkel «Snubler Norge inn i en algoritmisk velferdsdystopi?» (2022), adresserer hvordan KI-teknologi kan endre forholdet mellom borger og stat. Forfatterne viser til FN sin tidligere spesialrapportør for ekstrem fattigdom og menneskerettigheter Phillip Alston, som i 2019 uttalte at verden er på vei inn i en digital velferdsdystopi (Alston 2019, referert i Broomfield og Lintvedt 2020, 2). Offentlig forvaltning har i følge Broomfield og Lintvedt tilgang til store mengder data, og sammen med økende digitalisering kan det skape et press for å ta i bruk data på nye og inngripende måter. KI-modeller kan i større grad bli brukt til å avdekke svindel og juks, enn å forbedre folks liv, som kan føre oss nærmere en kontrollstat, og dermed endre forholdet mellom borger og stat (2). Selv om Broomfield og Lintvedt (2020) påpeker at de ikke anser at Norge er på vei mot en algoritmisk velferdsdystopi, fremmer artikkelen et viktig argument: dersom teknologi ikke brukes og utvikles klokt og etisk forsvarlig i Norge, kan forholdet mellom borger og stat endres, der borgere sin frihet vil reduseres.

En annen artikkel som tar for seg bekymringer ved bruk av KI innenfor norsk offentlig sektor er Ruetter og Broomfield (2019) sitt studie «Kunstig intelligens/data science: En kartlegging av status, utfordringer og behov i norsk offentlig sektor - første resultater», som er et samarbeid med Difi, NTNU og UIO. Studiet var en kartlegging av bruk og behov knyttet til bruken av KI innenfor offentlig sektor. I studiet utførte forfatterne en spørreundersøkelse av virksomheter i offentlig sektor, og dybdeintervju. Gjennom analysen ble det avdekket at få av informantene anså etikk som en utfordring innenfor sitt arbeid, noe forfatterne mener kan stamme fra uvitenhet rundt etiske dilemmaer. Så lenge løsninger tok for seg personvern og lovlighet, anså informantene løsningene som etisk (Broomfield og Reutter 2019). Selv om dette studiet er fem år gammelt innenfor et felt i stadig endring, reiser det viktige diskusjoner.

Antakelsen informantene hadde om at en løsning var etisk så lenge det var lovlig og personvernet var ivaretatt, åpner for spørsmål om hvordan etiske vurderinger oppfattes og fungerer i praksis.

I Leonora Bergsjø og Inga Strümke sitt debattinnlegg «Etter to år med nasjonal strategi for kunstig intelligens trengs opplæring og struktur» (2022), skriver forfatterne at selv om regjeringens mål om at KI-systemer «skal bygge på etiske prinsipper, respektere menneskerettighetene og demokratiet, ivareta den enkeltes integritet og personvern og være sikre» (Bergsjø og Strümke 2022) er en selvfølgelige, er det på samme tid ambisiøst. Forfatterne peker på hvordan det ikke alltid er en selvfølge at KI-systemer respekterer etiske prinsipper. I Norge bruker befolkningen digitale verktøy som følger «betenkelige» digitale praksiser. Eksempler presentert av forfatterne er Google og Facebook, som begge har praksiser som strider mot prinsippet om pålitelig og etisk forsvarlig KI. Bergsjø og Strümke mener at regjeringen må prioritere etisk digitalisering, men for at det skal kunne realiseres er det et sterkt behov for kompetanseheving innen feltene KI og digital etikk. Forfatterne mener også at vi må skape bedre strukturer for kontroll av systemer basert på KI (Bergsjø og Strümke 2022).

Behovet for etisk kontroll er også nevnt i Riegler, Lepperød og Røstads (2023) debattinnlegg «Norge som fanebærer for etikk i en uforutsigbar framtid», hvor forfatterne stiller spørsmålet: "Hvordan kan vi sørge for at KI-utviklingen er etisk og gagner hele samfunnet?". Om man ser på etisk bruk av KI på et internasjonalt nivå, redegjør forfatterne for hvordan de store teknologiselskapene kontrollerer KI-forskning og utvikling uten tilstrekkelig etisk tilsyn. Store selskaper tar ikke hensyn til data eller etikk og forskningsprioritetene deres er styrt av kommersielle interesser, som kan føre til etisk tvilsomme utfall, skriver forfatterne. "Uavhengig tilsyn og regulering av kommersiell KI er nødvendig for å sikre at teknologien er i tråd med menneskelige verdier og forhindre misbruk" (Riegler, Lepperød og Røstad 2023).

Et eksempel på et regelverk som har fungert godt i følge Riegler, Lepperød og Røstad (2023) er GDPR (General Data Protection Regulation), som er en forordning med regler som gjelder for alle EU/EØS land (Datatilsynet 2023). Forfatterne argumenterer for at Norge har en god posisjon for å lede utviklingen av KI-løsninger som er etisk rettet, vitenskapelig og

nyskapende, ettersom den skandinaviske samfunnsmodellen legger en god ramme for rettferdig, innovativ og gunstig KI-utvikling. For å sikre etisk rettet og ny KI forslår forfatterne følgende syv tiltak: 1) «Øke kunnskapen i samfunnet gjennom å innføre opplæring i ansvarlig bruk av KI i alle utdanninger», 2) «Etablere veiledning i å utvikle KI-programmer som tar rettferdige beslutninger», 3) «Utarbeide retningslinjer for transparens og åpenhet i KI-systemer», 4) «Ikke vente på innføringen av AI Act, men allerede nå etablere et veilednings- og tilsynsorgan for KI i Norge», 5) «Prioritere grunnforskning på KI som sikrer at vi forstår teknologien og besitter kompetansen i Norge som er nødvendig for å utvikle etisk og ansvarlig KI», 6) «Prioritere tverrfaglig samarbeid i utviklingen av KI, slik at vi bedre kan forstå alle nødvendige perspektiver», og 7) «Etablere et felles-nordisk samarbeid om utviklingen av etisk og ansvarlig KI» fordi Norge alene er lite, mens i Norden er det 27 millioner mennesker (Riegler, Lepperød og Røstad 2023).

4.1.3 Regulering: Hvilke lover skal sikre god bruk av KI i Norge?

I Norge har vi i dag ingen lover som er dedikert til bruk av KI. I Baste, Schultz og Osbergs (2023) artikkel «Mens vi venter på at EU skal regulere kunstig intelligens», kommer forfatterne med viktige poeng når det gjelder utviklingen innenfor regulering av KI i Norge. Selv om Norge ikke har en KI-lov på plass enda, redegjør Baste, Schultz og Osberg (2022, 17) for hvilke regler og lover vi har i dag til å sette rammer for bruk og utvikling. Helt grunnleggende er vi forpliktet gjennom grunnloven og menneskerettsloven. Andre lover, som likestillings- og diskrimineringsloven, forvaltningsloven og sektorspesifikk lovgivning, er også relevante for å sette rammer for utviklingen i dag (Baste, Schultz og Osberg 2023). Forfatterne peker på en stor usikkerhet hos ulike aktører knyttet til hvordan man skal anvende eksisterende lover når man ønsker å ta i bruk KI. Dette begrunner de med at dagens lover og regelverk mangler relevans, og det forfatterne beskriver som "spesifisiteten" som nødvendig for KI. En løsning kan ifølge forfatterne være å utvikle en egen KI-lov i Norge, en annen mulighet vil være å tolke og tilpasse eksisterende juridiske rammeverk så de kan adressere muligheter og utfordringer innenfor KI mer presist (Baste, Schultz og Osberg 2023, 17).

EUs KI-forordning kan sette rammer for norsk regulering. Et viktig bidrag til en mulig KI-lov er EU sin Artificial Intelligence Act, som er en ny forordning foreslått av EU-kommisjonen. Forordningen er verdens første forslag til et juridisk rammeverk som spesifikt regulerer KI

(Datatilsynet 2023; Lanestedt, Goodwin og Andersen 2023). I skrivende stund jobber EU med å ferdigstille KI-forordningen, som i stor grad vil påvirke bruk av KI på et internasjonalt nivå. Dette er viktig for Norge, fordi gjennom EØS-avtalen vil mulig EUs KI-forordning måtte gjøres til norsk lov. Dette medfører en rekke spørsmål rundt hvordan Norge skal tilpasse sine nasjonale regelverk (Baste, Schultz og Osberg 2023, 15). På denne måten vil regulering av KI-teknologi i Norge påvirkes av EU sitt arbeid (Baste, Schultz og Osberg 2023, 17). Forfatterne understreker også at Norge ikke er medlem av EU, og derfor har Norge hatt liten innflytelse på utformingen av KI-forordningen.

Riegler og Lysne (2024) mener at mens vi venter på at KI-forordningen implementeres, har Norge en unik mulighet til å utvikle strategiske satsninger og tilpasse nasjonale strategier for KI. Forfatterne presenterer tre tiltak som vil hjelpe Norge å ta ansvar: Første tiltak er å utvikle et nasjonalt datalager, som vil gi sikker tilgang til viktige og sensitive data. Andre tiltak er å få på plass en robust infrastruktur, for å gjøre det mulig å raskt trene, tilpasse og teste KI-modeller. Siste tiltak er å utvikle et evaluerings-rammeverk for å kunne evaluere KI-modeller. Om et rammeverk er basert på våre etiske og tekniske standarder og samfunnsverdier kan det bidra til å sikre at KI-systemer samsvarer med Norges behov (Riegler og Lysne 2024).

4.1.4 KI-milliarden: Regjeringens storsatsing

I september 2023 presenterte regjeringen en milliardsatsing på kunstig intelligens, hvor pengene skal gå til forskning på KI og digital teknologi de neste fem årene (Kunnskapsdepartementet og Statsministerens kontor 2023). Regjeringen ønsker at satsingen vil bidra til større innsikt rundt konsekvenser av teknologiutvikling i samfunnet, og bidra til å gi mer kunnskap til nye digitale teknologier, og muligheter for innovasjon i offentlig sektor og næringslivet. Forskningssatsingen har tre hovedspor: 1) forskning på konsekvenser av KI og annen digital teknologi. 2) Digitale teknologier som forskingsområde. 3) Forskning på digitale teknologier og innovasjon i offentlig sektor og næringsliv (Kunnskapsdepartementet, Statsministerens kontor 2023).

I en pressemelding utgitt av Kunnskapsdepartementet i mars 2024, skriver de at Norge vil i løpet av 2025 få inntil seks nye forskningssenter for KI, hvor statsråden oppfordrer til samarbeid mellom offentlig sektor, næringsliv og akademia. Hele 850 millioner av KI-

milliarden vil bli brukt på å finansiere de nye KI forskningssentrene (Kunnskapsdepartementet 2024). Forskningssentrene skal ta tak i problemstillinger som kunnskapsdepartementet mener vil kreve en tverrfaglig tilnærming som kobler forskning på innovasjon, teknologiutvikling og samfunnskonsekvenser.

4.2 Litteraturundersøkelse del 2: Internasjonalt nivå

KI-etikk er et forskningsfelt som har vokst fram på bakgrunn av bekymringer knyttet til partiske utfall innenfor automatisert beslutningstaking (Kazim og Koshiyama 2021). Et svar på de økende bekymringene har vært å ty til regulatoriske tiltak som retningslinjer og prinsipper for å kontrollere teknologien (Munn 2022). Denne delen av litteraturundersøkelsen vil gjennom forskning undersøke bias, med særlig fokus på hvordan det oppstår og hvilken betydning begrepet har. Den vil også belyse utviklernes innflytelse på systemene, undersøke hvordan KI-etiske prinsipper fungerer i praksis, redegjøre for alternative tilnærminger og til slutt beskrive EUs innflytelse på feltet.

4.2.1 Bias

Et sentralt spørsmål innen KI-etikk er: «Hvordan kan vi unngå bias?». Bias er en kritisk faktor innen KI-etikk, fordi det kan ha alvorlige konsekvenser når bias enten oppstår eller forsterkes gjennom automatisering (Kirkpatrick 2016). For å kunne ekskludere bias må man først forstå hvordan bruk av KI og algoritmer kan frembringe bias. Bias kan bli inkorporert i KI-systemer gjennom utvikling, men også gjennom interaksjon med de ferdigstilte KI-systemene (Kirkpatrick 2016; Howard og Borenstein 2017, 525). Kirkpatrick (2016) skriver at KI-modeller kan prosessere og behandle betydelig større datasett enn mennesker, der algoritmer trenes opp med et datasett som er samlet inn av maskiner, mennesker eller en kombinasjon av begge. Videre fatter algoritmen beslutninger på bakgrunn av beregninger basert på datagrunnlaget. Gjennom disse prosessene mener Kirkpatrick (2016) at det er mulig at KI-systemer fanger opp en eller annen form for systematisk bias eller menneskelig bias.

Før man kan kategorisere *bias* som et negativt fenomen, er det viktig å forstå at begrepet har ulike betydninger avhengig av forskningsfelt. For å undersøke forskjellige former for bias innen KI har jeg tatt utgangspunkt i to sentrale artikler: Danks og Londons (2017)

«Algorithmic Bias in Autonomous Systems» og Ferrer mfl. (2021) sin artikkel «Bias and Discrimination in AI: A Cross-Disciplinary Perspective». Disse artiklene gir en bred oversikt over ulike typer bias. Med utgangspunkt i deres analyser har jeg valgt å fokusere på fire typer bias som er særlig relevante for denne oppgavens problemstilling: *treningsdata-bias*, *prosesseringsbias*, *overføringsbias* og *tolkningsbias*.

Den første formen for bias jeg vil trekke fram fra artiklene er *treningsdata-bias*, som er en form for bias som ofte stammer fra avvik under utvikling og opplæring av algoritmen, eller fra datagrunnlaget brukt for å trene algoritmen (Danks og London 2017, 2). Dette kan skje når datasett brukt for å trene algoritmen inneholder eksisterende fordommer og/eller tidligere beslutninger, som i flere tilfeller kan reflektere i algoritmens beslutning (Ferrer mfl. 2021, 7). Dette er en form for bias som Ferrer mfl. (2021, 73) mener kan være skjult, fordi utviklere sjeldent offentliggjør hvilke data de bruker til å trene algoritmene sine. Derfor mener forfatterne at man ikke alltid er klar over at partiske data ble brukt til trening når vi bruker ferdigstilte KI-modeller.

Den andre formen for bias er *prosesseringsbias*. Bias kan bevisst introduseres når man trener en algoritme, noe som gjøres for å dempe eller kompensere for skjevheter i dataene (Ferrer mfl. 2021, 73). Når en algoritme i seg selv er partisk, kan man bruke det Danks og London (2017, 3) kaller en «skjev estimator». En estimator kan i følge EITCA (2023) forklares som en algoritme eller matematisk funksjon, som tar inn data og videre produserer et estimat av en funksjon eller målevariabel. Estimatorer er designet for å lære mønstre og sammenhenger for å forutsi presise estimeringer (EITCA 2023). Å velge en skjev estimator kan være et bevisst valg for å redusere andre former for bias, som vil gjøre algoritmen mer pålitelig og robust (Danks og London 2017, 3). Prosesseringsbias trenger ikke nødvendigvis å være negativt – det kan faktisk bidra til å opprettholde viktige moralske normer og prinsipper. Likevel har diskusjoner om bias ofte blitt hemmet av ulike tolkninger av begrepet. Noen ganger brukes det som et nøytralt beskrivende uttrykk, andre ganger med en negativ klang, noe som kan skape forvirring og vanskeliggjøre diskusjonen om hvordan og når man skal reagere på bias som oppstår med algoritmer (Danks og London 2017, 3). Dette er en viktig type bias å fremheve, ettersom mange ofte forbinder begrepet med noe utelukkende negativt. Ettersom jeg nå har vist at begrepet *bias* kan ha ulike betydninger, vil jeg understreke at jeg i resten av

oppgaven vil bruke bias i en negativ betydning. Dette fordi jeg refererer til urettferdige utfall som oppstår når KI-systemer introduserer eller forsterker historiske, sosiale og menneskelige fordommer i samfunnet gjennom automatisering, og ikke refererer til begrepets tekniske betydning.

De to første typene bias viser til tekniske aspekter ved algoritmer, som hvordan de er bygget og trent opp. De to siste: *overføringsbias* og *tolkningsbias*, handler derimot om hvordan algoritmene anvendes i praksis. *Overføringsbias* oppstår når en algoritme brukes på områder eller situasjoner den ikke er designet for (Ferrer mfl. 2021, 73), som kan føre til feilaktige resultater som ikke oppfyller juridiske, moralske eller statistiske standarder (Danks og London 2017, 4). Dette skyldes, ifølge Danks og London (2017, 4) ofte feil bruk av algoritmen heller enn selve algoritmens tekniske utforming. Forfatterne skriver at slike feil kan likevel føre til alvorlige konsekvenser, slik som bias ut fra en moralsk standard, hvor algoritmen produserer skjeve eller urettferdige resultater. Ved å trekke fram brukerfeil som en kilde til bias, understrekes det at ansvar for etiske utfordringer med KI ikke alltid ligger i teknologien selv, men også i hvordan den implementeres og brukes. Dette gir en nyanse til diskusjonen om ansvar i KI-systemer (Danks og London 2017, 4).

Den siste formen for bias, og kanskje den som er viktigst for oppgavens problemstilling er *tolkningsbias*. Denne formen ser nærmere på feiltolkning av algoritmers utdata eller funksjon av de som samhandler med algoritmen (Danks og London 2017, 4; Ferrer mfl. 2021, 73). *Tolkningsbias*, kan ifølge Danks og London (2017, 4) skyldes et misforhold mellom systemet og brukeren som anvender algoritmens utdata. Selv når algoritmen brukes innenfor riktig operasjonskontekst, kan feiltolkning likevel oppstå. Forfatterne skriver at en grunn for feiltolkning er at utviklere ikke kan spesifisere algoritmens eksakte semantiske innhold. Etter hvert som flere og flere arbeidsoppgaver blir delegert videre til det digitale, mener Danks og London at det vil det kunne bli en utfordring for de som bruker ferdigstilte KI-verktøy å tolke algoritmens beslutninger. Uten digital kunnskap og forståelse over hvordan det algoritmiske verktøyet de tar i bruk er bygget, kan det være utfordrende å evaluere algoritmens utfall (Danks og London 2017, 4).

Som illustrert over har *bias* begrepet ulike betydninger i forskjellige kontekster. Ferrer mfl. (2021) utfordrer i sin artikkel bias-begrepet og viser heller til *digital diskriminering* for å skille de negative og positive formene for bias. Forfatterne mener at all urettferdig behandling som stammer fra automatisert beslutningstaking bør anses som *digital diskriminering* og ikke *bias*. Dette begrunner forfatterne med at bias ikke alltid fører til diskriminering: «Bias betyr et avvik fra standarden, noen ganger nødvendig for å identifisere eksistensen av noen statistiske mønstre i dataene eller språket som brukes» (Ferrer mfl. 2021, 72, egen oversettelse). Innenfor ulike forskningsdisipliner har ordet bias ulike betydninger. I informatikk refererer bias til et avvik fra standarden, og det er ikke alltid et tydelig skille mellom diskriminering og bias. Bias fungerer som et viktig verktøy for å identifisere og klassifisere forskjeller, men det fører ikke nødvendigvis til diskriminering (Ferrer mfl. 2021, 72).

4.2.2 Konsekvenser av bias

Ferrer mfl. (2021) skriver at automatiserte systemer og algoritmer opererer i stor skala og kan påvirke en rekke fenomener som kan forårsake skade dersom de fatter urettferdige beslutninger. Ved bruk av KI-verktøy og systemer innen offentlig administrasjon forklarer Ferrer mfl. at mennesker potensielt kan settes i fare hvis noe går galt i prosessen. Forfatterne understreker at KI stadig tar over viktige beslutninger, som helseevalueringer og saksbehandling, noe som kan få kritiske konsekvenser for både enkeltpersoner og samfunn. Derfor mener forfatterne at automatiserte systemer kan fatte urettferdige beslutninger ved å kategorisere visse grupper mennesker systematisk dårligere gjennom automatiseringsprosessen. Ferrer mfl. identifiserer dette som et voksende problem, ettersom stadig flere beslutninger delegeres til automatiserte verktøy (Ferrer mfl. 2021, 72).

Et eksempel på hvordan bias kan ta form er hentet fra artikkelen «Gender Shades» (2018) skrevet av Buolamwini og Gebru, en artikkel som har fått massiv respons siden den ble utgitt. Artikkelen tar for seg rase- og kjønnsbias innenfor ansiktsgjenkjenningsverktøy, og avdekker hvordan algoritmer kan diskriminere på grunnlag av rase og kjønn. I artikkelen analyserer forfatterne tre kommersielle klassifiseringssystemer. Systemene de analyserer viste seg å ha en feilmargin på opp mot 34,7% for kvinner med mørk hudfarge, i motsetning til feilmarginen for menn med lys hudfarge som er på 0,8%. For å forstå bakgrunnen for disse feilmarginene, og hvorfor det er så drastiske forskjeller, analyserte forfatterne to datasett brukt for å utvikle

ansiktsgjenkjenningsverktøy. Analysen viste at 79,6% og 86,2% av dataene brukt var av mennesker med lys hudfarge (Buolamwini og Gebru 2018, 1). Dette er en form for *treningsdata-bias* som introdusert tidligere.

Norge har ikke vært like tidlig ute som andre land med å ta i bruk KI-systemer i offentlige organ. Derfor er det mye kunnskap å hente fra andre land sine feil ved bruk av teknologien. I USA har de utarbeidet en database for ansiktsgjenkjenning som er tatt i bruk av rettsvesenet, hvor 117 millioner amerikanere er representert. En forskningsundersøkelse basert på data hentet inn gjennom flere år, fra 100 ulike politistasjoner i USA, viser at mennesker med mørk hudfarge har høyere sannsynlighet for å bli analysert av ansiktsgjenkjenningsverktøy enn mennesker med andre etnisiteter (Garvie mfl. 2016, referert i Buolamwini og Gebru 2018, 2). Slike former for automatiserte systemer innenfor rettshåndheving kan utgjøre en trussel for siviles rettigheter når systemene feilidentifiserer på bakgrunn av farge, kjønn og rase (Klare Accountability 2012, referert i Buolamwini og Gebru 2018, 2). Et annet eksempel hentet fra bruk av KI-systemer innenfor amerikansk rettshåndhevelse er *COMPAS*-algoritmen. *Compas* ble utviklet til å fungere som et saksbehandlings- og beslutningsverktøy med mål om å bistå amerikanske domstoler i å vurdere tiltaltes sannsynlighet for å begå nye lovbrudd (Larson mfl. 2016). En analyse skrevet av Larson, Mattu, Kirchner og Anwin (2016) av denne algoritmen viser at mennesker med mørk hudfarge har større sjanse for å bli feilberegnet som høyrisiko enn mennesker med lys hudfarge (Larson mfl. 2016). «Gender Shades» og *Compas* er høyprofilerte eksempler på hvilke konsekvenser bruken av partiske systemer kan ha.

4.2.3 Menneskene bak systemene

Bias knyttet til rase og kjønn slik som i «Gender Shades» (Buolamwini og Gebru 2018) er et komplekst problem, men D'Ignazio og Klein (2020) mener at man i flere tilfeller kan knytte paralleller mellom hvordan bias oppstår og hvem som står bak KI-systemer. Menneskene som utvikler, designer og samler data til til KI-systemer er ofte en liten gruppe mennesker som tilhører dominerende grupper, som oftest hvite menn. En konsekvens av dette kan være at deres perspektiver på hvordan beslutninger skal tas videreføres inn i algoritmene de utvikler, og at andre perspektiver blir ekskludert. Dette er som oftest ikke en bevisst handling: D'Ignazio og Klein peker på at dominerende grupper ofte er uviten om andre perspektiver, noe som kan resultere i lite mangfold av perspektiver i automatiseringsprosessen (D'Ignazio

og Klein 2020, 28). Manglende mangfold hos utviklerne kan være en forklaring på hvorfor algoritmisk bias slik som i «Gender Shades» (2018) kan oppstå.

Uten mangfoldig representasjon i datasett brukt for å utvikle KI-systemer, kan skjevhetene leve videre og føre til urettferdighet som tar nye former. Når bias forekommer innenfor KI og automatisert beslutningstaking, har det en tendens til å forsterke og utvide eksisterende skjevheter. Samtidig legger slike systemer til rette for utvikling av nye klassifiseringer og kriterier, som igjen kan danne grunnlag for fremveksten av nye former for bias (Ntoutsi mfl. 2020, 2, Mehrabi mfl. 2021, 2). Spesialiserte KI-systemer stammer fra forskere og selskapene som utarbeider systemer for å videreføre sine modeller til offentligheten (Howard og Borenstein 2017). Når algoritmer utvikles, blir de tildelt hva Howard og Borenstein (2017) kaller for sannhetsetiketter. Når KI-systemer viderefører eksisterende bias, får forskere og selskaper som utvikler disse systemene makt til å forme og etablere nye globale sannheter (Howard og Borenstein 2017, 1524). Samlet fremhever D'Ignazio og Klein (2020) og Howard og Borenstein (2017) viktigheten av å se på menneskene bak programvaren. Det vil derfor være viktig for Norge under utvikling av nye KI-verktøy at vi ikke bare prioriterer mangfold i datasettene, men også mangfold i hvem som utvikler verktøyene.

Etter hvert som KI-teknologi tar en større rolle i våre økonomiske, sosiale og politiske liv, mener Mullaney (2021) at det overføres både bevisste og ubevisste verdier fra utviklerne til selve datasystemene. Forfatterne skriver at denne prosessen fører til at skjevheter, ulikheter og former for marginalisering overføres til et nytt domene. Når verdier integreres i teknologiske systemer, endrer de karakter; de blir mer varige og usynlige, som kan gjøre dem mindre tilgjengelige for menneskelig revisjon og dermed vanskeligere å analysere, vurdere og evaluere. På denne måten kan ulikheter knyttet til religion, klasse, rase og kjønn finne veien inn i teknologier som automatiserte beslutningsverktøy, og potensielt anta en ny og mer skadelig form (Mullaney 2021, 6).

4.2.4 Regulering og KI-etiske prinsipper

Kazim og Koshiyama (2021) fremhever regulering som en av de mest etablerte tilnærmingene for å adressere KI-etiske utfordringer. Siden Norge følger en prinsippbasert tilnærming som går under regulering, vil jeg i dette delkapittelet redegjøre for sentrale bekymringer knyttet til denne tilnærmingen.

De siste årene har det vokst opp en økende bevissthet rundt etiske spørsmål knyttet til design, utvikling, distribusjon og bruk av KI-systemer (Kazim og Koshiyama 2021). Dette har vekket spørsmål rundt *ansvarlighet*, *personvern*, *rettferdighet*, *tilgjengelighet*, *bærekraft* og *åpenhet*, som alle er prinsipper som har blitt mye analysert og diskutert innenfor akademisk litteratur (Jobin, Ienca og Vayena 2019). Gjennom en økt bevissthet og forskning har det oppstått en enighet av at lover og regulatoriske rammer ikke er tilstrekkelig for å beskytte mot skade (Jobin, Ienca og Vayena 2019, Munn 2022). Innenfor KI-etikk og spørsmål knyttet til hvordan vi skal unngå eller stoppe urettferdige utfall på bakgrunn av automatisert beslutningstaking, har et svar vært å ta i bruk retningslinjer, rammeverk og prinsipper (Morley mfl. 2021; Jobin, Ienca og Vayena 2019; Munn 2022; Hagendorff 2020). I dette delkapittelet har jeg belyst fire tekster som alle evaluerer retningslinjer og nytten av dem.

Retningslinjer kan være vanskelig å praktisere. I Morley, Ellhalal, Garcia, Kinsey, Mökander, Flordi (2021) sin artikkel «Ethics as a service: a pragmatic operationalisation of AI ethics», påpeker forfatterne at regjeringer og organisasjoner har utviklet «myke» styringsmekanismer i form av etiske retningslinjer, prinsipper, koder og politiske strategier. Forfatterne skriver at det har blitt tydelig de siste årene at abstrakte prinsipper ikke gir tilstrekkelig beskyttelse mot potensielle farer og skader knyttet til KI. En grunn til dette, ifølge forfatterne, er at KI-utøvere ikke har fått god nok veiledning for hvordan man skal designe og distribuere algoritmer innenfor de gitte etiske grensene (Morley mfl. 2021, 840) Dette fører til et betydelig gap mellom teori og praksis, som også andre forskere identifiserer (Morley mfl. 2021, Munn 2022, Jobin, Ienca og Vayena 2019). En annen grunn til at retningslinjer kan være vanskelig å praktisere er utformingen. Jobin, Ienga og Vayena (2019) i artikkelen «The global landscape of AI ethics guidelines», mener det er en global enighet om at KI-verktøy skal være etiske, men det er ikke en enighet rundt utformingen og innhold til ulike retningslinjer og rammeverk. Uenighetene er knyttet til hvilke etiske krav man skal stille, krav til teknisk

standard og til beste praksis for å oppnå en global etisk tilnærming (Jobin, Ienca og Vayena 2019).

Det er mange som stiller seg kritisk til bruken av retningslinjer, og en av disse er Munn (2022), som i sin artikkel «The uselessness of AI ethics», peker på KI-etiske prinsipper som ineffektive. Munn beskriver retningslinjer som «meningsløs», fordi de ofte er abstrakte og tvetydige. Retningslinjer inneholder ofte mange ulike prinsipper, Munn påpeker at de sjeldent gir spesifikke anbefalinger og adresserer heller ikke grunnleggende normative og politiske spenninger. Munn beskriver en mangel på konsensus rundt betydningen av nøkkelbegreper innenfor KI-etikk som et stort problem. Når det ikke er enighet rundt viktige begreper som *personvern* og *rettferdighet*, argumenterer han for at tolkningen av retningslinjer og prinsipper oppfattes på forskjellige måter som kan føre til inkompatible mål. Dette fører til at KI-etiske prinsipper blir usammenhengende og ineffektive i praksis (Munn 2022, 870).

I Jobin, Ienca og Vayena sin artikkel «The global landscape of AI ethics guidelines» (2019) analyserte forfatterne 84 ulike KI-etiske retningslinjer og demonstrerer en framvekst av en global konsensus rundt fem etiske kjerneprinsipper: *åpenhet*, *rettferdighet*, *ikke-ondskap*, *ansvar* og *personvern*. Utvelgelsen av retningslinjer bestod av ikke-juridiske dokumenter utstedt av ulike organisasjoner, der forfatterne analyserer eksisterende etiske prinsipper og retningslinjer for bruk av KI. Selv om forfatterne identifiserer en økt enighet om at disse prinsippene er viktige, påpeker Jobin, Ienca og Vayena at det likevel er betydelige forskjeller i hvordan de ulike prinsippene tolkes og implementeres på tvers av ulike kontekster. I sin analyse identifiserer forfatterne et gap mellom etiske prinsipper på høyt nivå, og deres praktiske bruk innenfor KI-utvikling. I likhet med Munn (2022), peker Jobin, Ienca og Vayena på KI-etiske prinsipper som abstrakte, og mener de mangler spesifikke anbefalinger. Ifølge Jobin, Ienca og Vayena er abstrakte prinsipper vanskelige å operasjonalisere på en effektiv måte innenfor virkelige omgivelser. Dette kan føre til et gap mellom teknologisk praksis og etiske idealer, noe som gjør KI-etiske prinsipper inkonsekvent og fragmentert.

Jobin, Ienca og Vayena argumenterer for en mer nyansert og kontekstsensitiv tilnærming til utvikling og implementering av KI-etiske retningslinjer. Som Munn (2022) også påpeker, ser de et behov for å adressere de sosio-tekniske utfordringene knyttet til KI-utvikling.

Forfatterne fremhever viktigheten av en helhetlig tilnærming til KI-etikk, som inkluderer en vurdering av undertrykkelsessystemer og sosiopolitiske dynamikker. Samtidig etterlyser de et mer spesifikt fokus på tekniske aspekter som styring, nøyaktighet og revisjon (Jobin, Ienca og Vayene 2019). Når KI-etiske prinsipper mangler politiske og sosiale kontekster, kan de bli *isolert*, som Munn (2022) argumenterer for. Isolerte prinsipper, ifølge Munn, fører til at prinsipper ikke klarer å adressere grunnleggende ulikheter og underliggende sosiale problemer som er med å forme teknologisk utvikling. På denne måten kan etiske, sosiale og moralske virkninger av KI bli oversett (Munn 2022, 870-871). Hagendorff (2020) sin artikkel «The ethics of AI ethics: An evaluation of guidelines», adresserer også de samme bekymringene, der han skriver at den «mannlige måten» å tenke på etiske problemer er reflektert i nesten alle etiske retningslinjer, der han refererer til KI-etiske prinsipper som: *ansvarlighet*, *personvern* og *rettferdighet*. Men at etiske bekymringer som *omsorg*, *pleie*, *velferd* og *sosialt ansvar*, derimot sjeldent blir adressert i KI-etiske retningslinjer mener (Hagendorff 2020, 103).

Hagendorff (2020, 114) argumenterer også for at etiske prinsipper og retningslinjer som er utarbeidet for å regulere bruk og utvikling av KI, ikke har innvirkning på menneskelig beslutningstaking innenfor maskinlæring og KI. Hagendorff skriver at KI-etikk på mange måter svikter, som kan skyldes mangel på en forsterkende mekanisme slik som konsekvenser for å bryte retningslinjene (114). Munn (2022, 871) argumenterer også for at KI-etiske prinsipper er det han beskriver som «tannløse», som han mener stammer fra mangel på håndhevingsmekanismer. De fleste utarbeidede rammeverk og prinsipper setter det Munn beskriver som «normative idealer», men mangler midler for å sikre overholdelse. Det kan bidra til å gjøre KI-etiske prinsipper om til et markedsføringsverktøy for bedrifter som man kan knytte til etikk-vasking (Munn 2022, 871-872). Dette kan tolkes som at KI-etikk har blitt markedsstrategi, og om man ser på KI-etikk i praksis, blir retningslinjer ofte ansett som fremmede eller et overskudd til tekniske bekymringer (Hagendorff 2020, 114). Hagendorff (2020, 115) understreker at flere anser retningslinjer som et uforpliktende rammeverk som er pålagt i institusjoner som står utenfor tekniske fellesskap. Ifølge Hagendorff (2020, 115) kan en mulig løsning på utfordringen med manglende kunnskap om bredere eller langsiktige samfunnsteknologiske konsekvenser være å fordele ansvaret mer systematisk. Dette kan bidra

til å motvirke utviklernes manglende følelse av ansvarlighet for de moralske implikasjonene av arbeidet deres.

4.2.5 Alternative tilnærminger til KI-etikk

I det forrige delkapittelet redegjorde jeg for sentrale bekymringer knyttet til prinsippbaserte tilnærminger innen KI-etikk. I dette delkapittelet vil jeg derimot utforske alternative tilnærminger som enten kan erstatte eller supplere KI-etiske retningslinjer og prinsipper. Den første tilnærmingen er av Kazim og Koshiyama (2021) beskrevet som "ethics by design" som forfatterne definerer som: "en forpliktelse til å bygge systemer etisk, i håp om at skade kan forebygges" (2021, 5, egen oversettelse). Forfatterne påpeker at det finnes flere tilnærminger for «ethics by design». Den første tilnærmingen er «co-design», som, ifølge Kazim og Koshiyama, refererer til tverrfaglig samarbeid under designprosesser. En fordel med denne tilnærmingen, er at KI-ingeniører ikke alltid er best posisjonert til å forstå etiske dimensjoner og virkninger av teknologien, derfor argumenterer forfatterne for at en tverrfaglig tilnærming er viktig for å inkludere flere perspektiver i designprosessen. En annen tilnærming til «ethics-by-design», er rettet mot å utvikle klare retningslinjer, prinsipper, lover og standarder for å strukturere og bedømme design. Denne tilnærmingen vil i følge Kazim og Koshiyama, være viktig for å motvirke mangelen på konsensus innen digital regulering, noe forfatterne mener gjør det vanskelig for ingeniører å etablere «beste praksis» i å omsette prinsipper til praksis. I tillegg til manglende konsensus, er det også utfordrende for teknologer å sette grenser mellom KI-etiske prinsipper, når de går på bekostning av hverandre. Vektlegging av et prinsipp framfor et annet må i følge Kazim og Koshiyama begrunnes og adressert innenfor ulike kontekster (Kazim og Koshiyama 2021, 5).

Innenfor den prosessbaserte tilnærmingen til KI-etikk oppstår spørsmål om styring. Kazim og Koshiyama (2021, 5) beskriver hvordan styring, med tanke på nye digitale teknologier, kan deles inn i to kategorier: *teknisk* og *ikke-teknisk* styring. "Ethics by design" faller under teknisk styring og omfatter prosesser og systemer som gjør teknologisk aktivitet ansvarlig og transparent. Dette innebærer å begrunne designvalg og sikre at systemene er tilgjengelige. Ikke-teknisk styring fokuserer derimot på systemer og prosesser for å opplyse beslutningstakere. Dette kan være i form av opplæring og utdanning, som forfatterne mener vil kreve kontinuerlig oppdatering fordi teknologisk utvikling går fort. Det vil også være

viktig å holde mennesker opplyst om hvordan automatiserte beslutninger brukes, samtidig som menneskerettigheter respekteres (Kazim og Koshiyama 2021, 5).

I Morley mfl. (2021), ser forfatterne nærmere på hvordan KI-etikk kan gjøres nyttig for KI-utøvere og reflekterer rundt spørsmål om hvorfor prinsipper og tekniske oversettelsesverktøy fortsatt er nødvendig selv om de er begrenset, og hvordan begrensningene kan overvinnes gjennom teoretisk forankring (241). Morley mfl. argumenterer for at KI-etikk kan dra nytte av å hente inspirasjon fra andre vellykkede former for anvendt etikk. Medisinsk- og forskningsetikk som begge kombinerer lovgivning, retningslinjer for etisk styring, praksis og prosedyrer, er former for anvendt etikk som, ifølge Morley mfl. (252), har klart å finne en god balanse mellom fleksibilitet og å være streng, og mellom sentralisert og delegert. Morley mfl. argumenterer for at - om man klarer å flytte fokuset på KI-etikk bort fra prinsipper og prosedyremessighet, kan KI-etikk være lettere å anvende for KI-utøvere, og om vi klarer å oppnå riktig balanse, kan et «pro-etisk» designarbeid lykkes (Morley mfl. 2021, 252).

Selv om Morley mfl. (2021) har tro på denne tilnærmingen, anerkjenner de også at den har begrensninger. Dette begrunnes med at man ikke kan kontrollere KI-systemer fullstendig gjennom teknisk design. Urettferdig KI er ikke kun et resultat av partiske datasett, KI-teknologi er også kompleks og uforutsigbar, forfatterne understreker at det vil være vanskelig å evaluere effekten av «pro-etisk» design før systemenes tas i bruk. På bakgrunn av dette identifiserer forfatterne et behov for regelmessig evaluering av KI-systemer, og peker på kvalitativ forskning og empirisk testing som en viktig kilde for å forstå ulempene av teknologien i detalj (Morley mfl. 2021, 252-253).

Bias og digital diskriminering har i følge Ferrer mfl. (2020), vært et tema for forskning innen en rekke ulike disipliner som sosiologi, juss, humaniora, medisin og informatikk. Selv med mye forskning rundt bias og digital diskriminering har ingen av disiplinene klart å løse problemet alene. Om man ser problemstillingen fra et teknisk perspektiv, mener forfatterne at informatikken ikke tar for seg beregningsmetoder for etiske eller sosiokulturelle kompleksiteter (72). Både evaluering og design av KI-systemer er forankret gjennom ulike perspektiver, mål og bekymringer. Å anta at det finnes en forhåndsdefinert vei gjennom disse

ulike perspektivene mener forfatterne er misvisende, Ferrer mfl. argumenterer for at det vil kreve tverrfaglig samarbeid for å finne gode løsninger for diskriminering innenfor KI (78).

4.2.6 EUs innflytelse

Målet med denne oppgaven er ikke å analysere og evaluere spesifikke retningslinjer, slik som Hagendorff (2020) har gjort. Likevel vil jeg trekke fram EU-kommisjonens etiske retningslinjer for pålitelig KI, ettersom Norges «Nasjonal strategi for kunstig intelligens» i stor grad bygger sin KI-etiske tilnærming ut ifra EU-kommisjonens retningslinjer (Kommunal- og moderniseringsdepartementet 2020). EUs retningslinjer er utviklet av en ekspertgruppe opprettet av EU-kommisjonen i 2018 (HLEG, 2019). Disse retningslinjene fremmer tre hovedprinsipper for KI:

> *1. det skal være lovlig, i samsvar med alle gjeldende lover og forskrifter;*
> *2. det bør være etisk, og sikre overholdelse av etiske prinsipper og verdier; og*
> *3. den skal være robust, både fra et teknisk og sosialt perspektiv, siden AI, selv med gode intensjoner systemer kan forårsake utilsiktet skade* (HLEG, 2019, 5).

Målet med retningslinjene, ifølge HLEG, er å fremme pålitelig KI. HLEG understreker at lovlighet, etikk og robusthet bør være oppfylt gjennom hele livssyklusen til KI-systemer. Selv om HLEG anser hver av disse komponentene som nødvendige, er det ikke tilstrekkelig å oppfylle bare ett av kravene for å oppnå pålitelig KI. Dersom det oppstår spenninger mellom dem, skriver HLEG at samfunnet bør samordne disse. Dokumentet fremhever også at, selv om lovlighet er ett av de tre hovedprinsippene, er lovlighet ikke eksplisitt behandlet, ettersom det primært fokuserer på veiledning knyttet til de to andre hovedprinsippene (HLEG 2019, 2).

Etter at EU-kommisjonenes retningslinjer (HLEG 2019) ble publisert, har flere kritisert dem. En av disse er Larsson (2020), som stiller seg kritisk til konstruksjonen av ekspertgruppen. Ekspertene, utvalgt av EU-kommisjonen til å delta i gruppen, bestod av forskere innenfor ulike forskningsfelt og representanter fra sivilsamfunnsorganisasjoner, men også representanter fra industri, slik som Google. Dette kan skape spenninger, ved at mennesker som representerer industri blir gitt for mye kontroll over regulatoriske spørsmål knyttet til bruk av KI. Ifølge Larsson (2020) har også mennesker innenfor ekspertgruppen kritisert

retningslinjene i etterkant, fordi de mener de bidrar til å muliggjøre etikk-vasking. Larsson (2020) mener at denne bekymringen stammer fra hvordan forbud mot visse bruksområder ble tonet ned, spesielt av representanter fra industri. En annen bekymring er hvordan juridiske spørsmål eksplisitt er utelatt fra retningslinjene (Larsson 2020).

Kapittel 5: Intervju

I denne delen av det empiriske grunnlaget presenterer jeg en tematisk analyse av datagrunnlaget fra intervjuene. Drøfting av funnene kommer i kapittel 6. Underoverskriftene i dette kapittelet er basert på informantenes uttalelser. Temaene er: KI-etikk: sett fra humanistiske og tekniske perspektiv, utfordringer knyttet til KI-etikk, kompetansebehov, norsk regulering av KI, EU sin påvirkning, utfordrende områder å implementere KI, forebygging av bias i norsk offentlig sektor og hvordan kan vi sikre etisk bruk av KI i Norge?

Humanister og teknologer ser på de samme problemene fra forskjellige perspektiv, og det ønsket jeg å få fram i denne oppgaven. Som presentert i Kapittel 3: Metode, har intervjuobjektene doktorgrad, fast ansettelse og forsker på temaer relevant for oppgavens problemstilling. Jeg utførte to intervjuer. Informant 1 har humanistisk bakgrunn, og forsker og underviser i digital etikk. Informant 2 har teknisk bakgrunn, og forsker og underviser i tekniske fag.

5.1 KI-etikk: sett fra humanistiske og tekniske perspektiv

For å kartlegge informantenes erfaring og posisjon innen feltet KI-etikk, og hvordan de anvender etikk i sitt arbeid, startet jeg begge intervjuene med å spørre om deres forhold til KI-etikk. Informant 1 forteller at hun har arbeidet med KI-etikk som forskningsfelt i en årrekke. Hun anser forskingsfeltet som enormt viktig når KI har fått en så stor samfunnspåvirkning. Informant 2 har et annet forhold til KI-etikk, med en teknisk bakgrunn. Hun forteller at hun har observert debattene rundt etikk, lovlighet, rettferdighet, skjevheter og andre områder, men hun vil ikke plassere seg selv som en ekspert innen KI-etikk. Hun forteller at hun kjenner til feltet og bruker det i arbeidet sitt, men at hennes arbeid er mer teknisk rettet. Selv om informant 2 ikke anser seg selv som en ekspert på området, er hun en verdifull kilde til innsikt i hvordan tekniske disipliner forholder seg til KI-etiske bekymringer.

For å få en tydeligere forståelse av informantenes KI-etiske posisjon, spurte jeg dem om deres tanker og meninger rundt etikk-begrepet. Informant 1 forteller at når man snakker om etikk, så går det fort over i det juridiske:

> *...Som vil si det går veldig fort til regulering og spørsmål om personvern og så blandes det inn i den litt større paraplyen som heter trustworthy KI, som egentlig handler om å være lovlig, sikker og etisk. Det første jeg alltid etterspør er et rent fokus på det etiske, ellers så faller det fort bort i diskusjonen rundt å regulere eller å ikke regulere*

Informant 2 som har en teknisk bakgrunn forteller at for hun handler etikk om valg, fordi hun mener designvalg har etiske konsekvenser. Informant 2 påpeker at hun ikke er tilhenger av skillet mellom verdi og fakta. Hun mener etikk har blitt en integrert del av teknologien, hvor man gjør etiske valg enten man er klar over det eller ikke. Informant 2 sier at det derfor er viktig at vi stiller spørsmål under utvikling:

> *Hva trenges for å realisere teknologien, hvilke ressurser, materielle ressurser, hvilken kraft behov har den, og hvilke konsekvenser får bruk for enkelt person, samfunn og verden?*

Informant 2 fortsetter med å beskrive hvilke etiske vurderinger hun møter på gjennom sitt arbeid:

> *Når noen bestemmer at man skal prioritere teknisk performance, så er det objektivt ikke et verdiladet valg i seg selv, men det blir verdiladet fordi man kan ende med å velge bort etiske vurderinger som – hvem gagner på dette, hvem taper på dette, hvem vinner på det og hvilke risiko introduserer man for både intenderte og uintenderte konsekvenser?*

Jeg spurte begge informantene hvordan deres forskningsfelt skiller seg fra andre innen konteksten av KI-etikk. Informant 1 forteller at kritisk tenking og å stille spørsmål er en stor del av det vi arbeider med innenfor humaniora. Dette mener hun preger norsk kultur, og at vi i Norge har det hun kaller en «kritisk kultur», som ikke bare peker på humaniora. I Norge har vi kultur for at elever kan si nei til læreren, barna til foreldrene og ansatte til ledere. Informant 1 understreker at man ikke skal langt ut i Europa før det er en helt annen kultur. Hun forteller at «en stor del av humaniora er å ta vare på dannelsesperspektiv som handler om å tenke

bredt, som involverer mange ulike former for kunnskap og ikke-spesialiserte utfordringer innenfor teknologi». Hun sier at innenfor humaniora så jobber vi med store idealer slik som ideen om menneskeverd og hva det betyr i praksis, politisk, pedagogisk og historisk».

Informant 1 forteller at de humanistiske verdiene som beskrevet over er noe hun mener man ser mindre av innenfor andre fag slik som de tekniske, hvor man ser at perspektiver som likeverd og likestilling ikke tematiseres. Informant 1 stiller spørsmål rundt om teknologer får nok opplæring til å føle seg kompetente når etiske spørsmål kommer opp.

> *Nå setter jeg det på spissen. Det er teknologer som snakker mye om det etiske ansvaret, og så er det de som sier: kan vi ikke bare få lov til å gjøre vår greie? - så kan dere gjøre det andre. Og jeg ønsker meg jo flere som tørr å gå inn i de tverrfaglige samtalene og si at, her ser jeg noe, er det et etisk problem? – det må vi snakke om. Ikke sant, de ser noe annet en du og jeg klarer å oppdage av etiske utfordringer.*

Informant 1 forteller at det er en forestilling om at etikere kun snakker på et veldig abstrakt nivå, og om det normative. Her mener hun vi må passe på å ikke gjøre det, men heller snakke om anvendt etikk, som handler om det praktiske, og de løsningsorienterte diskusjonene, sånn at vi kan gi noe til samtalen.

Informant 2 ser også et behov for mer tverrfaglig samarbeid hvor teknologene kan lære mer om etiske perspektiver og humanistene mer om tekniske perspektiver. Hun mener etiske vurderinger er viktige og nødvendige diskusjoner innenfor teknologimiljøer. Hun forteller videre at mye har skjedd innenfor teknologimiljøene de siste årene, for eksempel ved at det er økt oppmerksomhet på risikoene ved KI og økt fokus på hvordan man skal programmere for å unngå ulike former for bias. Videre forteller informant 2 at det for ikke-teknologer kan bli lett å stå på utsiden og finne de dramatiske skandaleeksemplene på hvor galt det kan gå. For informant 2 er dette viktig og riktig, men det har begrenset verdi for teknologene. Hun beskriver det som viktig at de tverrfaglige samarbeidene skjer tidligere i prosessene for å få større effekt. Informant 2 understreker at det krever mye å sette seg inn i en teknologi, og å forstå hva man ser og hører når man jobber tett på teknologer:

> *...Men ikke sant, sånn å lære om algoritmer, de trenes opp med belønning og straff – for å si det med metaforer. Og klarer man å se når disse belønningsmekanismene dannes, så er det kjempesentralt for å si hvilke verdier det er som programmeres inn her. Og senere får det konsekvenser. Men ideelt sett så burde humanister samfunnsvitere og alle inn så tidlig som mulig i utviklingen tenker jeg. Ikke bare når det er ferdig å slippes ut.*

Jeg spurte informant 2 om mennesker som jobber innenfor informatikk og tekniske fag har et annet forhold til etikk når det er de som utvikler modellene. Hun forteller at hun skulle ønske at flere teknologer tenker at når man gjør teknologidesign, så er man en faktor av «moralske agenter»:

> *Designvalg har konsekvenser - om man ser på bygninger og tilgjengelighet så har vi lover som regulerer dette. Velger du å bygge trappetrinn, velger du å ekskludere noen mennesker, velger du å designe den digitale Interface på denne måten, så er det et valg om å inkludere eller ekskludere, Så jeg tenker vi er moralske agenter og det er for lav bevissthet om det.*

Informant 2 fortsetter med å fortelle at selv om teknologene har et stort ansvar, så er det ikke teknologene som tar alle valgene. De setter teknologien ut i verden, og videre må også organisasjonene som tar i bruk teknologien være ansvarlige og evaluere hva teknologien egner seg for, evaluere risikobilde og hvordan de skal ta i bruk teknologien på en fornuftig måte.

5.2 Utfordringer knyttet til KI-etikk

Et av spørsmålene jeg stilte begge informantene, er hva de syns er de største utfordringene innenfor KI-etikk. Informant 2 forteller:

> *Det er jo at det er vanskelig. Det er lett å liste opp etiske retningslinjer. Du skal være rettferdig, du skal være ikke-diskriminerende, du skal ditt og datt, også er det okei - hvordan finner vi ut om vi er det eller ikke? Hvordan måler vi, hvordan evaluerer vi teknologien?*

Informant 2 forteller at i tilfeller med mye data kan man gjøre en statistisk solid vurdering. Men om man ser på etnisitet, er det vanskelig å evaluere skjevhet i forhold til dette når man ikke systematisk har registrert etnisitet i datagrunnlaget.

> *Man kan for eksempel rigge en studie som går tre år fram i tid og systematisk registrere det, men det krever finansiering, samtykke og samarbeidspartnere for å om tre år kunne gjøre den vurderingen.*

Hun forteller at det er ofte «tungt, dyrt og krevende» å evaluere og måle for å finne skjevheter og bestemme hva man skal gjøre med dem.

Informant 1 derimot, forteller at den største utfordringen innfor KI-etikk, er at vi ikke snakker nok om det etiske:

> *Den største utfordringen vil jeg si er at vi ikke snakker om det, altså til enhver tid så fokuserer vi på teknologi som er mulig, og hvilke grenser lovene setter. Og så er det en stor debatt om hvor mye innovasjon og hvor mye risikovilje man skal kunne ha i det mulighetsrommet som er da mellom den teknologien som er tilgjengelig og den som er ferdig regulert og hvor det er ganske fastlåst hva man kan gjøre. Så det å ha en ordentlig debatt som ikke bare handler om loven, men faktisk det etiske. Vi er ikke kommet lengre enn at det er første skritt.*

Informant 1 fremhever at spørsmål om etikk ofte går over i det juridiske. Hun mener at i debatter rundt KI er det viktig å ikke kun tenke regulering, men også på de etiske perspektivene hvor vi må stille spørsmål til hvilket samfunn vi ønsker oss på lang og kort sikt:

> *Min erfaring er at når man snakker om etikk, så går det veldig fort over i jussen, som vil si det går veldig fort til regulering og spørsmål om personvern, også blandes det inn i den litt større paraplyen som heter trustworthy KI, som egentlig handler om å være lovlig, sikker og etisk. Det første jeg alltid etterspør er et rent fokus på det etiske, ellers så faller det fort bort i diskusjonen rundt å regulere eller å ikke regulere.*

Informant 1 understreker at vi trenger et tydelig skille mellom det etiske og det juridiske. Hun forteller at mange temaer faller innenfor en gråsone. Hun viser til begrepet «privacy» og forklarer at innenfor dette prinsippet er det aspekter som omhandler både etikk og juss, og det samme gjelder for andre prinsipper som for eksempel «ansvarlighet» og «ikke-diskriminering». Derfor mener informant 1 at det er viktig at man viser at man snakker om den større diskrimineringen, også det som ikke er lovfestet, for å tydelig demonstrere at man snakker om det etiske.

5.3 Kompetansebehov

Et av spørsmålene jeg spurte begge informanter om, er om de anser det som viktig at mennesker som skal ta i bruk KI-verktøy i offentlig sektor har digital kompetanse, og kompetanse knyttet til algoritmer og hvordan verktøy utvikles.

Informant 1 forteller at en viss forståelse er viktig, slik som å forstå at språkmodeller er statistiske modeller og ikke «sannhetsguder»:

> *Altså, å forstå på hvilket nivå teknologien beveger seg, tenker jeg er avgjørende for å kunne vurdere hvilken god bruk man kan gjøre med dem. Sånn at i en eller annen form for AI-literacy, for å bruke et godt norsk ord, eller ethical-AI-literecy, som da er de etiske aspektene ved den kunnskapen du trenger om KI, for å kunne bruke det klokt. Jeg tenker at det må inn i alle utdanninger og videreutdanninger rett og slett, for det*

> *er helt avgjørende etter hvert som disse verktøyene nå brukes mer og mer, særlig i offentlig sektor, men også over alt.*

Informant 1 forteller også at det er viktig at alle som går ut av skolen framover trenger å vite hvordan vi kan anvende KI på en positiv måte. Derfor er det viktig at lærerstudenter får god opplæring, og at elever skal bli bedre rustet enn dagens unge som ikke har hatt opplæring i denne tematikken. Informant 1 etterlyser økt politisk fokus på KI, og mener at den manglende viljen blant politikere kan skyldes en mangel på kompetanse.:

> *Hvis man ser på det politiske programmene i dag, så ser man jo ikke mye spor av digitalisering. Jeg tenker jo at Rødt burde formulere noe annet enn Høyre når det gjelder digitalisering. At politikken og de politiske linjene burde synes i hvordan de tenker om digitalisering. Det ser vi ikke i dag - og det tror jeg mangler litt kompetanse og etterspørsel fra velgere, for at de skal formulere tydelig hva det betyr.*

Informant 2 ser også et behov for økt kunnskap om KI i Norge, og mener at flere burde ta utdanning og interessere seg for det. Med erfaring fra helsesektoren forteller hun at det i mange år har blitt forsøkt å få mer teknologikunnskap inn i grunnutdanning til leger og sykepleiere, noe som har vært vanskelig, for «da må noe annet ut, og hva skal det være?», sier informant 2.

På spørsmål om det trengs kunnskap på hvordan teknologien vi bruker er utviklet og funger i praksis, forteller informant 2 at hun er splittet fordi vi bruker mye teknologi vi ikke forstår, og likevel bruker vi den på en måte vi opplever som trygg. Hun mener det er stor forskjell på verktøy, og at forskjellige bruksområder kan kreve forskjellige tilnærminger. Om man for eksempel bruker støtteverktøy for tekst, er det ikke så viktig å vite hvordan det fungerer, men det er viktig å danne seg en erfaringsbasert forståelse til å kunne evaluere om programmet oppsummerer bra og finner riktige kilder. Men om man bruker KI for å få innsikt i store datamengder og videre bruker den innsikten til beslutningsstøtte, mener informant 2 det vil være positivt at brukere av slike verktøy har innsikt i hvordan verktøyet fungerer.

Informant 2 har jobbet innenfor helsesektoren med KI-verktøy laget for å avdekke sykdom. Jeg spurte om hun syns det er viktig at leger som tar i bruk verktøyene har innsikt i datagrunnlag, og har forståelse for om hvorfor modellen gjør som den gjør:

> *Jeg tenker at de som utvikler og godkjenner det må skjønne det, og så er det ikke så farlig, om de tar ansvar og sier nå har vi sjekket, organisasjonen bestemmer at dette bruker vi, så kan de som kommer etter bruke, og slappe litt mer av tenker jeg.*

Dette begrunner hun med at innenfor organisasjoner kan man ha personer som er ansvarlig for å følge med og å ta grep ved avvik og problemer. Hun syntes det er for mye å forvente at enkeltbrukere skal gjøre hele jobben, så det er viktigere at organisasjoner som helhet har systemer som følger teknologien tett. Informant 2 mener at å øke kunnskapsnivået om KI generelt i befolkningen er en bra ting, men samtidig kan man ikke forvente at alle skal kunne alt. Hun forteller at dette har med teknologien sin modenhet å gjøre.

I land som Finland, har de utviklet digitale kurs i KI, jeg spurte informant 1 om dette er en god tilnærming. Hun forteller:

> *Ja det vi jobber med nå, det er AI literecy som er i disse kursene.* det *krever ganske mye innsikt til teknologien som jeg tenker ikke er realistisk. Ikke sant, så vi trenger enklere versjoner for bestemødre og lærere med dårlig tid ikke sant. Som kan gi den aller viktigste informasjonen for å ivareta både sikkerhet og personvern, men også disse demokratiske dilemmaene som vi kommer i når teknologien blir brukt på en uklok måte. Så et kurs som er mer tilpasset hadde jeg ønsket meg. Som ikke har så høy terskel for å gå inn, som ikke er så omfattende.*

5.4 Norsk regulering av KI

For å sikre etisk og god bruk av teknologien, er det ofte et svar å ty til retningslinjer, noe som også er relevant i norsk kontekst. Derfor spurte jeg begge informanter om hva de syntes om dagens retningslinjer, prinsipper og regulering av teknologien i Norge.

Informant 2 forteller:

> *Jeg tror det er veldig vanskelig for la oss si en liten, eller små og mellom små bedrifter som ikke har en forskningsavdeling, eller som ikke har ressurser til å sette seg inn i det. Så det er vanskelig å vite hvor man skal begynne. Så vi kunne hatt noe mye mer konkret. Jeg tenker de er gode de prinsippene, men det er mer å gå fra det nivået her oppe, ned til - oi hva gjør vi da? - som er utfordringen. Det er ikke noe galt med prinsippene i og for seg.*

Informant 1 mener at de som koder og de som bestiller løsninger ikke klarer å omsette de store begrepene som brukes i retningslinjer til praksis. Et eksempel hun viser til er hvordan begrepene *rettferdighet* og *ikke-diskriminering* er vanskelig for mange å oversette til praksis. Så selv om retningslinjene som finnes i dag er gode i seg selv, så er de vanskelig å anvende for dem som skal ta konkrete valg i den enkelte virksomheten, forteller informant 1.

Jeg spurte informant 2 om hun merker et gap mellom teori og praksis, der hun forteller:

> *Jeg merker jo at det er lettere å snakke om hva vi vil ha, enn å få det til.*

Likevel mener hun at det er positive ting som skjer, spesielt når det gjelder konkretisering. Informant 2 viser til Datatilsynet sin regulatoriske sandkasse som belyser problemstillinger, dilemma og løsninger, og publiserer rapporter for andre som lurer på det samme. Dette syns informant 2 er en god modell, fordi det skjer læring på tvers. Denne måten å jobbe konkret på har hun stor tro på. Informant 2 syns det skjer mye bra, men at det også kunne skjedd enda mer.

Informant 2 forteller videre at hun tidligere har vært skeptisk til det hun beskriver som «fluffy» og høynivå-ord, men at hun senere har forstått verdien av dem:

> *Det gir likevel en retning, så om du går imot disse så kan man påtale det at - vi har sagt at vi skal ha fairness, og du leverer ikke det, og dette er ikke greit, så de kan bety noe likevel.*

Hun mener høynivå-ord har virkning ved at når man har blitt enig om noe, så må politikken forholde seg til det og jobbe innenfor de rammene. Hun ser et behov for å konkretisere mer, men anser likevel prinsippene som gode, fordi de gir et godt utgangspunkt:

> *Og det å være enig i at dette er våre grunnleggende verdier, at dette er viktig, dette er mer viktig enn å være et innovasjonsvennlig samfunn som blir rikest i verden, det er et viktig valg.*

Da jeg spurte Informant 1 om hva hun syns om dagens retningslinjer og regulering etterlyser hun bedre veivisere:

> *Altså, den er ikke god nok. Et problem handler om denne oversettelsesproblematikken. Det er ikke gode nok retningslinjer for hva slags prosesser man kan ha, etterhvert så begynner det å bli noen gode veivisere innenfor det lovlige aspektene, ikke sant – likestillings- og diskriminerings ombudet kom jo med en veiviser nå før jul, som omhandler ikke diskriminering, og det har noe med privacy by design det er jo det vi snakker om non-discrimination by design, men retningslinjer for ethics by design det har vi ennå ikke, og det tenker jeg vi trenger å få på plass for å kunne lage løsninger som er etisk sikre nok. Så vi har retningslinjer det er supert, men anvendelsen må vi jobbe med.*

Jeg spurte informant 1 om hvordan man kan sikre at teknologien blir brukt godt. Hun påpeker at mange ønsker mer regulering og tilsyn. Samtidig fremhever hun at en mulig konsekvens av dette kan være at det stenger enkelte dører. Det ideelle, ifølge henne, ville vært om bransjen selv kunne regulere seg, for å unngå unødvendige tiltak:

Det vi har sett foreløpig er jo at bransjen ikke gjør det. Ikke viser vilje til å ta de grepene som er nødvendig. Det gjør jo etikken sant, etikken gjør ikke ting lettere, billigere eller raskere. Det betyr ikke at det blir kjempedyrt og dårlig. Man må prioritere noe, og det gjelder jo innenfor bærekraft og likestilling og mange områder at det har noen omkostninger å prioritere noe. Men viljen til å prioritere etikken, den vil også koste, det må virksomheter være villig til å ta. Jeg tenker jo at det kan det offentlige som vi snakker om bare bestemme, politikere kan si at vi tar den etiske riktige løsningen, og ikke den raskeste og den billigste hver gang. Det har vi oljemilliarder til å gjøre hvis politikerne vil. Så her kan vi virkelig gå foran som et godt eksempel hvis det er politisk vilje.

På oppfølgingsspørsmål om hvilke konsekvenser det har om man ikke følger retningslinjer i Norge, forteller informant 1:

På det etiske så vil det ikke være konsekvenser utover omdømmet fordi det er det lovlige, du kan ha tilsyn hvor du kan ha bøter eller andre type tiltak. Jeg vet jo også at oppfølgingen er veldig dårlig. For eksempel på hvor mange virksomheter som deler kundenes data, og til dels sett sensitive data. Det har det vært noen saker i det siste med blant annet et apotek på nett der det kommer fram at virksomheten ikke følger opp det som er loven, så selv der er det lite tilsyn og lite oppfølging. Så vi er ganske langt unna at det blir oppfølging på at mangfold og inkludering og ikke-diskriminering og disse enda løsere begrepene som brukes og som omhandler etikk.

5.5 EU sin påvirkning

Norges KI-etiske tilnærming er i stor grad utarbeidet i samsvar med EU sine retningslinjer og prinsipper for pålitelig KI, og vil i nær framtid bli påvirket av EU sin KI-forordning, som beskrevet i 4.1.3. Derfor ble begge informanter spurt om deres tanker rundt tematikken.

Informant 1 beskriver det som «for svakt» at Norge henter sine KI-etiske prinsipper fra EU. Hun begrunner dette med at det er viktig at man i Norge tenker på «den solidariske samfunnsmodellen som preger norsk velferdsmodell» i møte med KI:

> *Jeg syns det er for svakt at vi i Norge ikke tenker på hvordan den solidariske samfunnsmodellen som preger norsk velferdsmodell ser ut med KI. Det vi ofte ser i teknologien er at den blir brukt til det motsatte av det vi prøver å få til som bred humanistisk tradisjon, hvor alle skal med, og at alle er like mye verdt.*

Informant 1 forteller videre at «systemer som algoritmer legger til rette for, er noe av det motsatte av det vi prøver på med mangfold og inkludering»:

> *Alle disse systemene som algoritmene legger til rette for, er noe av det motsatte av det vi prøver på med mangfold og inkludering. Og i retningslinjene til EU så er jo det flere etiske punkter der som er veldig bra, men hva betyr de - for at vi skal tilsvare den samfunnsmodellen som vi ønsker, det tok ikke regjeringen diskusjonen på i det dokumentet hvert fall – Altså i sin nasjonale strategi for KI.*

Tidligere i samtalen spurte jeg informant 1, om hvilke lovlige kriterier som gjelder for aktører som utvikler teknologi. Hun viste til EU's sine retningslinjer for pålitelig KI og forteller:

> *For eksempel i ansvarlig KI konseptet hvor du har lovlig, etisk og sikker. Hvordan kan du ta vare på den etiske biten også? Du får diskusjonen internt i virksomheten som er i tråd med vårt mandat, og våre retningslinjer eller Norges. Er du i opplæringssektoren så er det jo noe om hvilke verdier skolen eller utdanningen skal bygge på. Synes det i løsningen? - Og hvis det ikke gjør det, så kan man ikke si at den er etisk forsvarlig, selv om den er lovlig og sikker.*

Informant 2 er mindre kritisk til at vi henter etiske prinsipper fra EU og forteller at hun har tenkt at det er positivt. Hun forteller at teknologiverden de siste årene har vært dominert av store spillere som ikke forholder seg til demokratiske spilleregler og peker på at techgigantene og plattformselskapene gjør som de vil:

> *..Så EU er en god ting, fordi de har en aktiv teknologipolitikk. Og dette var, altså AI ACT som vi også skal implementere, er jo også et produktsikkerhetsrammeverk som egentlig jeg tenker - det er ikke en dum måte å gjøre det på. Altså den risikobaserte, ja. Jeg tenker det er passende detaljeringsnivå eller tilgangsnivå. Du kan ikke være for teknologispesifikk, det blir så fort utdatert. Så jeg ser ikke så mye problemer med at vi følger EU.*

Jeg spurte begge informanter om hva de syntes om utformingen av ekspertgruppen (HLEG) som står bak EU sine retningslinjer for pålitelig KI, ettersom flere representanter fra industri – som Google, var en del av utvalget. Informant 1 mener at det har stor påvirkning:

> *Selvfølgelig får det jo påvirkning, det var mange som snakket om etikkvasking, mener jeg. Og når man ser hvor lite detaljert og hvor høyprofilert det er, så kan man jo absolutt tenke at her er det noen som har tenkt at vi lager noe som er så bredt at vi kan si at vi gjør det, men som ikke kan oversettes til en praksis som utfordrer oss.*

På oppfølgingsspørsmål om hvordan vi kan unngå denne fellen i Norge, sier informant 1:

> *Altså, politisk vilje. Vi har pengene, vi har høy digital kompetanse, høy infrastruktur. Og vi har en samfunnsmodell som er veldig opptatt av solidaritet og rettferdighet i mange betydninger. Så skal noen få det til, så bør vi gjøre det. Så bør vi selvfølgelig finne løsninger som kan eksporteres så vi også kan hjelpe andre da.*

Informant 2 er litt mer splittet til kommersielle aktører sin påvirkning i utformingen av EUs retningslinjer, og mener vi også må gi rom til de store aktørene når vi utarbeider retningslinjer for bruk av teknologien:

> *Jo, det er klart det er påvirkning fra teknologisiden, klart det er det. Jeg vet ikke balansen, men jeg mener begge deler er viktig. Man kan ikke tyne teknologisiden heller de kommersielle aktørene helt heller, de må få lov til å finne et marked, det må være mulig for dem å levere produkter og tjene penger nok til å overleve og vokse og videreutvikle.*

5.6 Utfordrende områder å implementere KI

Jeg spurte begge informanter om hvor de tror det vil være ekstra utfordrende å ta i bruk KI innenfor offentlig sektor.

Informant 1 beskriver det som utfordrende å ta i bruk teknologien innenfor områder som involverer sårbare grupper eller mennesker i vanskelige livssituasjoner. Hun mener at skolen kan være en utfordrende plass å ta i bruk KI. Det er variasjoner i praksis og kompetanse fra kommune til kommune, og man ser at mye ansvar er plassert langt nede i systemet. Informant 1 forteller at det også er utfordringer innen helse og velferd, men særlig i utdanningssystemet, hvor man tar i bruk blant annet digital læringsanalyse for å tilpasse undervisning. Her mener hun det er viktig at vi ikke lar teknologien forsterke fordommer, som den ofte gjør.

Informant 2 ser også bekymringer knyttet til implementering av KI i kommuner:

> *Jeg tror kommunene generelt er veldig presset, mange er små og de har ikke ressurser til at noen skal sette seg inn i dette og lære dette, jeg ser noen større kommuner jobber litt mer bevisst med det.*

Videre forteller informant 2 at Kommunes sentralforbund KS (kommunesektorens organisasjon) må ta en mer aktiv rolle i å formidle hvilke verktøy de har testet og hvor verktøyene egner seg godt, og etterlyser et større samarbeid på kommunenivå. Informant 2 trekker også fram helse som krevende fordi det er en todeling mellom spesialisthelsetjenesten og primærhelsetjenesten. Hun understreker at helse er en av sektorene som vil merke presset om effektivisering, noe som kan drive sektoren til å i større grad ta i bruk KI-verktøy.

Jeg spurte informant 2 om hvordan hun mener implementering av KI-teknologi i offentlig sektor fungerer i praksis. Hun forteller om digitaliseringsministeren sin uttalelse om at 80 prosent i offentlig sektor skal ta i bruk KI innen de neste to årene:

> *Til hvilken grad er et spørsmål, hva mener man med KI? - men også at hun sier det har ikke så mye å si om hun ikke får innarbeidet det i si Finansdepartementet, Justisdepartementet, fisker, at de også skal fortelle sine underliggende etater at nå skal dere gjennomføre det som hun sier. For hun har ikke styring over dem. Det ligger andre steder.*

Informant 2 forteller at Digitaliseringsdirektoratet skal være rådgivende innenfor digitalisering, men de har ikke styringsmyndighet nedover.

5.7 Forebygging av bias i norsk offentlig sektor

Ettersom KI i stor fart implementeres i norsk offentlig sektor, spurte jeg begge informanter om hva de mener man må være oppmerksom på for å unngå bias og urettferdige utfall.

Informant 2 jobber mye med å avdekke bias i sitt arbeid hvor hun analyserer data og datasett brukt for å utvikle KI-modeller. Hun jobber med modellbygging og prosesser, og analyserer hvorvidt man tar i bruk riktige rammer, grenser, monitoreringer og kvalitetssystemer rundt modellene. En av biasene hun trekker fram, er at hun ser en skjevhet i om innbyggere klarer å bruke verktøyene vi lager. Hun forteller at offentlige tjenester skal være universelle og tilgjengelige for alle, i motsetning til kommersielle tjenester. En annen bekymring for informant 2 er at datagrunnlag som finnes og brukes for å utvikle KI ikke alltid er representativt for befolkningen:

> *Du vet mer om noen grupper enn om andre, og dermed så kan man lage løsninger for det området og de gruppene man kjenner behovet til. Det er litt generelt kanskje for all teknologiutvikling.*

Informant 1 mener at diskusjoner rundt bias ofte er farget av tanken om at vi kan unngå bias. Hun sier det er urealistisk at maskiner ikke har en form for bias. Maskiner lærer og må ta valg, og i den prosessen ligger det valg av skjevheter. Ifølge informant 1 bør diskusjoner rundt bias ledes av hvordan vi skal unngå uønskede former for bias og partiskhet:

> *Vi må hele tiden sjekke – hvilke bias er det vi tar med oss i den løsningen, for vi klarer ikke å unngå det. Vi er nødt til å passe på at de løsningene vi har er etisk akseptable til valgene vi har gjort, og her ser man særlig de store språkmodellene at det flytter med seg forestillinger som er i vår verden, som har masse fordommer i seg, som forflytter seg inn i teknologien på måter som vi ikke klarer å forutse. Så det er derfor det med testing og evaluering og videre oppfølging av teknologien i hele livssyklusen med etisk risikovurdering, er avgjørende for å få etisk forsvarlige løsninger.*

I tråd med semi-strukturerte intervjuers åpning for spontane samtaler, spurte jeg informant 1 om hvordan sosiale skjevheter blir videreført inn i datasystem i Norge. Informant 1 forteller at i Norge har vi en forestilling om at det er få sosiale skjevheter i samfunnet, men at de forskjellene vi har, risikerer å bli økt når vi tar i bruk KI, spesielt innenfor skolen. Et eksempel hun viser til er forskning gjort i Danmark som viste at når elever brukte Google til å løse skoleoppgaver, ble søkene deres farget av hva de har søkt på før:

> *...Så ser du da at de som har søkt mye på kunnskapskilder typisk da barn med foreldre med høyere utdanning får bedre svar, skriver bedre oppgaver, får bedre karakterer og får bedre sjanser til å selv få en høyere utdanning. Ikke sant, men de som har foreldre med lav utdanning og har typisk flere underholdningssøk, så kommer flere underholdningsrelaterte svar. Så kan dette føre til dårligere karakterer og dårligere fremtidsutsikter.*

På denne måten kan sosiale skjevheter videreføres gjennom teknologien. Informant 1 mener vi er like sårbare som alle andre samfunn, selv om vi har mye likestilling mellom blant annet kjønn. «Likevel er det viktig å huske på at vi tar i bruk modeller fra andre land som har mindre likestilling enn Norge, noe som utfordrer noen av de humanistiske grunnverdiene som menneskeverd, likeverd, kritisk tenking og demokrati».

På spørsmål om skjevhet (bias) vil bli en utfordring når teknologien tas i bruk her i Norge, svarer informant 1 at biaser kan komme gjennom import av løsningene. Mange datasystem blir kjøpt fra utenlandske og ofte globale aktører og tilpasses før de blir tatt i bruk. På denne måten blir løsninger som er utviklet av en viss gruppe privilegerte mennesker, som ikke reflekterer det mangfoldet vi har eller ønsker oss i samfunnet.

Informant 2 viser også bekymring for kommersielle verktøy. Hun forteller at vi har lite kontroll over kommersielle verktøy siden bedriftene ikke publiserer kodene, men i firma som bygger sine egne analysemodeller, vet man hva de gjør og har mye mer kontroll. Her mener hun at både data-scientists og domene-eksperter må jobbe sammen:

> *Da er det ikke nok med en av gruppene, da må vi jobbe tverrfaglig for å få gode modeller som virker i den konteksten.*

Jeg spurte informant 2 om hva hun tenker om spenningspunktet mellom personvern og frihet, hun forteller at hver enkelte må ta sitt valg:

> *Ja, hver enkelt person tar jo sitt valg, vi pleier å kalle det convenience. Man er villig til å ofre mye privacy for convinience. Hvis du sier ja til posisjonstjenester på mobilen så får du dette og dette, det er det valget hver enkelt må ta.*

På et samfunnsnivå sett fra «social-contract-theory», har Informant 2 i sitt arbeid analysert hva innbyggere aksepterer eller forventer fra myndigheter:

> *Hva syns jeg er greit, for det er innbyggere, de er innbyggere i en stat, vi får velferdstjenester, så det er den relasjonen. Hva vil du gi for å få noe tilbake. Den balansen har på en måte vært samfunnsbestemt, hvor ligger den. Så får du krisesituasjoner som pandemi – så nå må vi flytte på den der grensedragningen mellom det private og det kollektive. For nå betyr det kollektive mer, nå må vi akseptere at vi blir mer inngripende. Så jeg tenker det er litt sånn politisk filosofi spørsmål - hvordan balanseres dette?*

5.8 Hvordan kan vi sikre etisk bruk av KI i Norge?

Jeg spurte begge informanter om løsninger på hvordan vi skal sikre etisk bruk av KI i Norge, og om tanker de mener er viktige i denne prosessen.

Informant 1 forteller at de viktigste skrittene mot en etisk implementering av KI i norsk offentlig sektor, er et høyere fokus på det etiske:

> *Vi trenger veivisere som fokuserer på etikk, som virksomheter som vil ta i bruk KI kan bruke, sånn at de ikke blir sittende med det gapet som er i dag mellom retningslinjer og sjekklister og den praktiske hverdagen de skal lage løsninger for. Så rett og slett bedre prosessverktøy.*

Informant 1 identifiserer også tre områder hun anser som viktig, som hun mener skiller seg ut og trenger høyere fokus. 1) Man må se på «ethics by design», som ser på hvordan teknologien oppfører seg. 2) Det er viktig å jobbe med «ethics for designers», en aktørfokusert tilnærming som skal sikre høy etisk kompetanse hos de som utvikler KI-modeller og for de som har ansvar for implementering av verktøy. Dette punktet mener informant 1 er viktig for alle yrkesgrupper som skal jobbe med teknologi, og ikke bare for teknologene. 3) Man må lage prosesser som ikke kun bruker retningslinjer, men som også involverer relevante interessenter («stakeholders»). Informant 1 mener det er viktig at brukere involveres tidlig nok for å bidra til å få etisk gode prosesser. «Dette må tydeligere inn i flere utdanninger – både for teknologer, sykepleiere og alle som skal jobbe med teknologi»

Informant 1 forteller også at hun har stor tro på etisk risikovurdering:

> *Det jeg jobber med og som jeg tror veldig på, det er etisk risikovurdering. Den faktiske diskusjonen om å vurdere risikoen ved de løsningene fra et etisk perspektiv. Altså ikke bare hvor mye kostnad blir det og så er det sikkert nok og er det lovlig, men hva gjør denne gjør denne løsningen med samfunnet på kort og på lang sikt. Og at man, det tror jeg ikke noen av oss kan gjøre alene, så det må gjøres tverrfaglig. Man må ha inn mange til å være med i den debatten, og det i en liten virksomhet trenger det*

ikke å være så mange, men det at å prioritere å gå den runden rundt etisk risikovurdering, det håper jeg flere og flere vil.

På spørsmål om hvordan vi skal sikre etisk bruk av KI i Norge, velger informant 2 å svare på et organisasjonsnivå. Hun forteller at om en organisasjon ønsker å ta i bruk KI, må man først stille spørsmål til hvorfor:

Første spørsmål er – hvorfor burde vi det, hva er formålet, hva vil vi oppnå det. Hvorfor skal vi det og hva av våre oppgaver kan løses bedre? Ha et formål- ikke bare gjør det fordi alle gjør det. Så å finne et formål, og så sette seg inn i hva er handlingsrommet, hva kan vi gjøre her, hvordan kan dette bidra positivt.

Når man har bestemt seg for å ta i bruk teknologien, forteller informant 2 at det følger en utprøvingsfase for å evaluere datagrunnlag og hvilken algoritme som er best egnet. I denne prosessen kan man oppdage at datagrunnlaget ikke er godt nok, der en løsning kan være å importere eller kjøpe mer data for å kunne stole på algoritmen vi utvikler. Her mener informant 2 at det er viktig med dialog med domeneeksperter for å evaluere teknologien. Når teknologien fungerer, så må man se på hvilke konsekvenser modellen har for arbeidsorganisering, hvilke oppgaver den skal overta, når den skal brukes, og hvilke ressurser skal spares, ifølge informant 2.

Informant 2 forteller videre om to ulike tilnærminger for å sikre etisk bruk: Den første vil være å utvikle strengere lovkrav – «det er forbudt å gjøre dette og du er påbudt å ha dette på plass før du skal ta i bruk KI». Hun foreslår en strengere tilnærming hvor man iverksetter en lov framfor bare retningslinjer. Den andre tilnærmingen er å tenke at de som utvikler teknologien må ha en høyere etisk bevissthet: en slags sertifiseringsordning for teknologer, sånn som for leger som har en profesjonsutdannelse og en lisens som du ikke får om du ikke støtter opp om verdisystemet i legestanden. Om man velger å følge den siste tilnærmingen, viser informant 2 til at det vil bli vanskelig å sette begrensinger for hvem som har lov til å utvikle teknologi. Hun understreker at teknologi er åpen kunnskap, i den form at man kan laste ned en artikkel som beskriver en algoritme og sette den i drift. Legene sin vellykkede

etikktilnærming stammer fra en monopolsituasjon gjennom 100 år, og et bevisst arbeid, som hun mener er vanskelig å få til i praksis:

> *Men en kombinasjon av fisk og gulerot tenker jeg er bra. Dette er prinsippene dette er. Høyrisikosystemer blir ulovlig, systemer med skadepotensiale vil vi ikke ha, og at tilsynene er aktive og får ressurser nok til å følge opp både Datatilsynet og det nye AI-tilsynet. Hvis det er en forskjell jeg ser gjennom Norge og Danmark, er at vi har et veldig pro-aktivt solid Datatilsyn som går ut og sier nei, her gjør dere feil, dette er ikke lov, dere får bot. Det oppdrar på en måte hele sektoren, ikke bare de som blir kontrollert, men også alle som hører om det.*

Jeg spurte informant 2 om hun mener det er en god strategi å prøve seg fram når det gjelder å regulere teknologien. Informant 2 forteller:

> *Ja, i denne sammenhengen så vet vi for lite til å være veldig bastante med en gang. Og det er så overraskende hva som skjer på teknologifronten. Jeg har fulgt med på dette en stund, men så plutselig så i høsten for 2 år siden kom ChatGPT med et bang. Det skjer dramatiske skifter så fort, og man må vite hva man vil, man skal ikke vingle, men man må kunne tilpasse ting.*

Informant 2 forteller også at offentlig sektor har en spesiell posisjon:

> *De har lovhjemmel for å hente ganske mye data og ganske private og personlige data fra innbyggere. Men jeg opplever at de er sitt ansvar bevisst, på at her kan vi ikke gjøre hva som helst. At de forholder seg til lovgivning og sånn. Men det er klart, det er jo dette dilemmaet som noen har skrevet om, som jeg har sett. Hvordan skal jeg uttrykke det? Altså, når man har så mye data, så kan vi gjøre dette, at vi gjør noe mer enn det man har gjort før. Det er mulig med mer overvåking og mer innsikt enn det man har hatt før. Og det må man jo på en måte være bevisst på og hvor langt er det riktig å gå i å bruke denne informasjonen.*

Kapittel 6: Diskusjon

Som Landestedt, Goodwin og Andersen (2023) påpeker, er den norske statsforvaltninger på vei inn i en «innovativ periode». Som beskrevet i kapittel 1 og 4, kan KI-teknologi gi flere positive effekter for offentlig administrasjon (Lanestedt, Goodwin og Andersen 2023), som den norske regjeringen ønsker å utnytte for å skape produktive, effektive og brukerrettede tjenester (Kommunal- og moderniseringsdepartementet 2020, 5). Selv om KI-teknologi kan ha positiv effekt i offentlig sektor, stiller mange seg kritisk til teknologien fordi den utgjør en betydelig risiko for utviklere, brukere, samfunn og menneskeheten, med lav forklaringsevne, fare for bias, ivaretakelse av datasikkerhet og personvern og andre etiske bekymringer (Siau og Wang 2020). Derfor er det viktig at offentlig sektor reflekter over etiske, partiske og samfunnsmessige implikasjoner av KI teknologi (Lanestedt, Goodwin og Andersen 2023).

Selv om KI kan gi fordeler innenfor offentlig sektor, må vi ikke glemme de etiske vurderingene i denne prosessen, og stille spørsmål til hvordan teknologien vi utvikler og bruker gagner hele samfunnet (Reutter, referert i Muhaisen 2024). Denne oppgaven har så langt sett teknologien gjennom en kritisk linse formet av humanistiske verdier, som setter menneskets beste i fokus. Dette verdigrunnlaget reflekteres også i diskusjonen, som ikke har som mål å redegjøre for teknologiens fordeler, men heller å diskutere punkter som utfordrer etiske hensyn i diskusjoner rundt implementeringen av KI i norsk offentlig sektor, og redegjøre for punkter som kan bidra til ivaretakelse av etiske hensyn. I dette kapittelet diskuterer jeg funnene fra den tematiske analysen av det empiriske materialet, sett i lys av oppgavens teoretiske bakgrunn, for å besvare problemstillingen:

> *Hvordan kan vi sikre at etiske hensyn står sentralt i implementeringen av kunstig intelligens i norsk offentlig sektor?*

Denne diskusjonen tar for seg KI-etiske prinsipper og deres funksjon og virkning, hvordan KI etikk fungerer i praksis i Norge, samt EU sin påvirkning og menneskene bak systemene. I diskusjonens siste kapitler redegjør jeg for to viktige tilnærminger som kan bidra til å bevare det etiske: økt digital kompetanse og tverrfaglig samarbeid.

6.1 KI-etiske prinsipper er vanskelig å oversette til praksis

I den nasjonale strategien for KI tydeliggjør regjeringen et ønske om å opprettholde et høyt tillitsnivå gjennom implementeringen av KI (Kommunal- og moderniseringsdepartementet 2020). For å sikre tillit til hvordan KI-teknologi utvikles, brukes og implementeres i norsk offentlig sektor, er det avgjørende at etiske hensyn står sentralt, og at regjeringens KI-etiske tilnærming legger til rette for etisk utvikling, bruk og implementering av KI. Regjeringens nasjonale KI-strategi understreker at norsk KI skal «bygge på etiske prinsipper», «respektere menneskerettighetene og demokratiet», og «være ansvarlig, robust, gjennomsiktig og ta hensyn til personvern» (Kommunal- og moderniseringsdepartementet 2020, 56). Etiske prinsipper som *ansvarlighet*, *robusthet*, *gjennomsiktighet* og *personvern* er gjentakende i internasjonale retningslinjer, inkludert EU-kommisjonens retningslinjer for pålitelig KI (HLEG 2019), som regjeringens KI-etiske tilnærming bygger på og i internasjonal forskning (Jobin, Ienca og Vayana 2019; Munn 2022; Kazim og Koshiyama 2018). Siden regjeringens etiske prinsipper er basert på et internasjonalt rammeverk, slik forskningsartiklene jeg har referert til viser, er funnene fra litteraturundersøkelsen også relevante for en norsk kontekst. I de neste delkapitlene i diskusjonen har jeg diskutert forskjellige bekymringer knyttet til regjeringens prinsippbaserte tilnærming.

Ifølge Jobin, Ienca og Vayena (2019) har det vokst fram en enighet for hvilke prinsipper som bør inkluderes i retningslinjer for KI, der *åpenhet*, *rettferdighet*, *ikke-ondskap*, *ansvar* og *personvern* er de prinsippene som oftest gjentok seg i de 84 KI-etiske retningslinjene forfatterne analyserte. Selv om dette viser en enighet om hvilke prinsipper som bør inkluderes, påpeker forfatterne at det ikke finnes en felles forståelse av hvordan disse prinsippene skal tolkes og implementeres. En av grunnene til at prinsippene er utfordrende å tolke, er deres abstrakte og tvetydige natur (Jobin, Ienca og Vayena 2019, Munn 2022, Hagendorff 2020). Uten en felles enighet viser internasjonal forskning at KI-etiske prinsipper er vanskelig å overføre til praksis.

Når vi ikke har en felles enighet om hva *åpenhet*, *rettferdighet*, *ikke-ondskap*, *ansvar*, *personvern* og andre KI-etiske prinsipper «faktisk» betyr, kan de tolkes ulikt av forskjellige aktører, som vil føre til variasjoner i KI-etisk praksis og hvordan prinsippene implementeres (Jobin, Ienca og Vayena 2019). Som informant 2 forteller, har bedrifter ulikt ressursgrunnlag,

der det er vanskelig for mindre bedrifter med mindre ressurser å sette seg inn i prinsipper og retningslinjer. Informant 1 støtter også opp under argumentet om at praktiseringen er utfordrende, der hun spesielt peker på begrepet *ansvarlighet* som noe som er vanskelig å oversette til praksis. Dette forsterker at utfordringer identifisert i internasjonal forskning, er relevant for den norske konteksten, der man også i Norge ser et gap mellom KI-etiske prinsipper og praksis.

Ulike algoritmer og KI-modeller bruker forskjellige teknikker, prosesser og data, og brukes til forskjellige formål (Kommunal- og moderniseringsdepartementet 2020). Med store variasjoner i utvikling og bruk av teknologien vil det være vanskelig å forholde seg til et abstrakt prinsipp, når konteksten man skal bruke modellen i, ikke er adressert i retningslinjer og KI-etiske prinsipper. Jobin, Ienca og Vayena (2019) etterlyser en mer kontekst-sensitiv tilnærming for hvordan man utvikler og implementerer KI-etiske retningslinjer. I kapittel fem «Ansvarlig og pålitelig kunstig intelligens» i nasjonal strategi, er ikke denne bekymringen redegjort for (Kommunal- og moderniseringsdepartementet 2020). For bedrifter med lite ressurser, som informant 2 belyser, kan det bli ekstra utfordrende å navigere seg i og implementere KI-etiske prinsipper, ikke bare fordi prinsippene er vanskelig å tolke og oversette til praksis, men også fordi det ikke er adressert hvordan de skal ivaretas i ulike kontekster.

KI-etiske prinsipper er ikke kun utfordrende for de som skal implementere og ta i bruk teknologien, men også for de som utvikler den. Kazim og Koshiyama (2021) påpeker at uklarhet og manglende enighet innenfor digital regulering gjør det vanskelig for teknologene å etablere «beste praksis» for å omsette KI-etiske prinsipper til ingeniørpraksis. Denne påstanden forsterkes av informant 1, som forteller at de som utvikler og bestiller løsninger ikke klarer å omsette de store begrepene i KI-etiske prinsipper til praksis. Informant 2, som representerer det tekniske, sier at hun har observert at det er lettere å snakke om hva man vil ha – enn å få det til. En forklaring på dette kan være informant 1 sin observasjon av at KI-etiske prinsipper ofte er så generelle at man kan hevde å ha fulgt dem uten at de nødvendigvis kan oversettes til en praksis som utfordrer oss. Sett i lys av Kazim og Koshiyama (2021) sin påstand om at det er vanskelig å etablere «beste praksis» for teknologer, så legger ikke dagens retningslinjer til rette for en god og etisk praktisering, fordi de er for generelle – eller abstrakt

og tvetydig som Munn (2022) argumenterer for. Totalt sett kan man si at gapet mellom teori og praksis av KI-etiske prinsipper påvirker flere ledd i teknologiutvikling og bruk, fordi det er vanskelig å videreføre dem til praksis både for de som bestiller, utvikler, implementerer og bruker teknologien. En tilnærming som kan bidra teknologer i å etablere «beste praksis», er tilnærmingen «ethics-by-design» presentert av Kazim og Koshiyama (2021). Tilnærmingen innebærer å utvikle klare retningslinjer, prinsipper, lover og standarder for å strukturere og bedømme design. En tilnærming informant 1 også beskriver som viktig å få på plass, for å kunne lage løsninger som er etisk sikre.

Når man ikke har klare «regler» for hvordan KI-etiske prinsipper skal overføres til praksis, kan mangel på enighet føre til at aktører anser en løsning som for eksempel *rettferdig*, så lenge den oppfyller en ofte abstrakt definisjon av rettferdighet og opererer innenfor lovens rammer. Som informant 1 påpeker, er en løsning ikke nødvendigvis etisk bare fordi den er lovlig. Broomfield og Reutter (2019) bemerker også at mange ser på løsninger som etiske så lenge de er lovlige og personvernet er ivaretatt. Uten klare definisjoner og konkretiserte prinsipper og retningslinjer, gir det rom for ulik tolkning som kan føre til en mindre konsekvent implementering og variasjon i etisk praksis på tvers av sektorer. Dette understreker behovet for en mer kontekstsensitiv tilnærming til KI-etikk, da et enkelt sett med retningslinjer ikke er tilstrekkelig for å imøtekomme de varierte behovene til forskjellige aktørene i offentlig sektor.

For å tette dagens gap mellom teori og praksis, mener informant 2 at Datatilsynets regulatoriske sandkasse kan belyse problemstillinger, dilemmaer og foreslå løsninger for andre som har spørsmål knyttet til bruk og implementering av KI. Derfor framstår Datatilsynet som en viktig ressurs i å bidra til å tette gapet mellom teori og praksis når norske bedrifter ønsker å ta i bruk KI-teknologi, fordi de kan konkretisere KI-prinsipper i praksis på tvers av sektorer, som er en form for prosessverktøy som informant 1 etterlyser mer av. Totalt sett er det er et betydelig gap mellom teori og praksis av KI-prinsipper i Norge, noe som ikke sikrer at det etiske står sentralt gjennom implementeringen av KI i norsk offentlig sektor.

6.2 KI-etiske prinsipper og norske kjerneprinsipper

I dette delkapittelet vil jeg diskutere hvordan ivaretakelse av ett KI-etisk prinsipp kan gå på bekostning av et annet, og hvordan norske kjerneverdier utfordres av de KI-etiske prinsippene. Kazim og Koshiyama (2021) skriver at vektlegging av prinsipper som *transparens,* kan gå på bekostning av *personvernet,* fordi prinsippene motsier hverandre. I den nasjonale strategien for KI erkjenner regjeringen at det kan være utfordrende å oppfylle alle de syv KI-etiske prinsippene samtidig, da det kan oppstå spenninger mellom dem som krever avveininger. Regjeringen påpeker at dersom en etisk akseptabel avveining ikke kan identifiseres, bør utvikling, utbredelse og bruk av løsningen «ikke fortsette i den samme formen». Samtidig unnlater strategien å gi konkrete anbefalinger om alternative tilnærminger. Når det ikke er utviklet klare linjer for hvor man skal sette grensen mellom prinsipper, vil det være vanskelig for utviklere og brukere å bestemme hvilket av de som skal prioriteres.

Regjeringen redegjør i nasjonal strategi at *åpenhet* (transparens) er et viktig norsk kjerneprinsipp, der borgere skal ha rett på å vite hvordan beslutninger som påvirker dem fattes. Regjeringen skriver også at vi i Norge har et grunnleggende personvernsprinsipp om *dataminimering*, som betyr at man skal begrense mengden innsamlede personvernsopplysninger til det som er nødvendig for formålet (Kommunal- og moderniseringsdepartementet 2020). I Datatilsynets (2022) sluttrapport for NAVs sandkasseprosjekt ble det fremhevet som en sentral bekymring at tiltak for å motvirke og avdekke diskriminering krever nye metoder, som innebærer behandling av flere og nye typer personopplysninger. Dette strider mot regjeringens sitt prinsipp om dataminimering (Kommunal- og moderniseringsdepartementet 2020), og er et konkret eksempel på utfordringer offentlige aktører kan møte på når de skal ta i bruk KI.

Når teknologien blir brukt på ulikt grunnlag og til forskjellige formål, vil det bli vanskelig for offentlige aktører og ivareta regjeringens prinsipp om *dataminimering* og *åpenhet* uten klare regler for hvor grensen skal settes. Som informant 2 påpeker, har offentlig sektor en særstilling, ettersom de har lovhjemmel til å benytte private og personlige data fra innbyggerne. Informant 2 forteller at offentlig sektor er bevisst i sitt ansvar og forholder seg til gjeldende lovgivning. Likevel beskriver hun et spenningspunkt i at mer data kan føre til mer overvåkning og innsikt enn før, som åpner for diskusjoner om hvor langt det er riktig å gå

i å bruke personlig data. Regjeringen (Kommunal- og moderniserinsdepartementet 2020) skriver i nasjonal strategi at det vil være viktig at det offentlige gir tydelig informasjon om hvordan KI-systemer fungerer. Dette innebærer å redegjøre for hvilke data som behandles, og hvilke formål de tjener. For å realisere *transparens* i tråd med norske verdier skriver regjeringen at det er også viktig at KI-baserte beslutningssystemer kan forklares. I tilfeller hvor det er viktig å kunne forklare hvordan algoritmiske beslutninger fattes, skriver regjeringen at et alternativ kan være å velge «andre tilnærminger» enn dyplæring. Regjeringen foreslår kun «andre tilnærminger» uten å gi anbefalinger på hvilke tilnærminger de mener kan være relevant. Selv om det er positivt at regjeringen redegjør for spenningspunktet mellom *åpenhet* og *personvern*, mangler det en tydelig diskusjon på hvor brukere og utviklere skal sette grensen mellom prinsippene. Som diskusjonen tidligere har beskrevet, er det behov for mer kontekst-spesifikk veiledning for ivaretakelse av KI-etiske prinsipper (Jobin, Ienca og Vayena 2019). Å foreslå «andre tilnærminger» uten konkrete anbefalinger er ikke tilstrekkelig for å bevare det regjeringen beskriver som viktige norske kjerneprinsipper, spesielt når det er et gap mellom teori og praksis som forskningen min har vist.

En annen bekymring belyst i det empiriske grunnlaget, er hvordan regjeringens KI-etiske tilnærming i stor grad er basert på EUs retningslinjer for pålitelig KI (HLEG 2019). Å adoptere et KI-etisk rammeverk som ikke tar hensyn til den norske konteksten, kan ha konsekvenser. Selv om EUs retningslinjer også omfatter prinsippene om *åpenhet* og *personvern*, adresserer rammeverket disse prinsippene på et annet grunnlag. I Norge er åpenhet forankret i et verdigrunnlag som er unikt for den norske konteksten, som ikke reflekteres i EUs rammeverk. Informant 1 stiller spørsmål ved hvordan EUs rammeverk harmonerer med den samfunnsmodellen Norge ønsker å opprettholde. Hun påpeker at denne diskusjonen ble oversett i regjeringens nasjonale strategi for KI (Kommunal- og moderniseringsdepartementet 2020). Vi kan ikke forvente at et internasjonalt rammeverk ivaretar norske verdier, men om vi tar i bruk EUs retningslinjer som de er, vil det kunne oppstå et spenningspunkt mellom norske kjerneverdier og KI-etiske prinsipper. Å følge EU sitt internasjonale KI-etiske rammeverk (HLEG 2019) som det er i dag, kan dermed føre til at norske verdier blir oversett, og viktige grunnverdier i samfunns- og velferdsmodellen blir nedprioritert gjennom automatisering. Som informant 2 forteller er det viktigere å prioritere våre grunnleggende verdier, enn å være et innovasjonsvennlig samfunn som blir rikest i

verden. KI-etiske retningslinjer i Norge bør gjenspeile de verdiene som preger norsk offentlig sektor og samfunn, for å støtte opp om velferdsmodellens solidariske fundament, mener informant 1. Om en KI-modell ikke støtter dette verdigrunnlaget, så kan man heller ikke kategorisere den som etisk, mener informant 1. Det empiriske grunnlaget fra intervjuene viser at regjeringen burde prioritere å evaluere sin KI-etiske prinsippbaserte tilnærming, for å sikre at etiske hensyn står sentralt. Status er i dag at de KI-etiske retningslinjene Norge bruker, ikke tar hensyn til politiske, sosiale og økonomiske kontekster eller offentlighetens forventninger til åpenhet og rettferdighet. Derfor kan tilliten regjeringen setter høyt, skades. En potensiell tilnærming kan være å utarbeide retningslinjer spesifikt for transparens og åpenhet i norske KI-systemer, som foreslått av Riegler, Lepperød og Røstad (2023).

6.3 Internasjonal påvirkning

I diskusjonen om hvordan internasjonale forhold påvirker norsk utvikling og implementering av KI, er EU sin påvirkning høyst relevant fordi Norge følger EUs KI-etiske prinsipper (Kommunal- og moderniseringsdepartementet 2020, HLEG 2019). Et viktig funn i det empiriske grunnlaget som belyser hvorfor det er viktig å diskutere EUs påvirkning, er Larsson (2020) som stiller seg kritisk til konstruksjonen av ekspertgruppen som utviklet EUs retningslinjer - fordi representanter fra sivilsamfunnsorganisasjoner og industri var med å utvikle retningslinjene. Larsson (2020) mener at dette kan bidra til å skape spenninger, når representanter fra industri blir gitt for mye kontroll over regulatoriske spørsmål. Informant 1 deler også denne bekymringen og peker på «etikkvasking» som en konsekvens. Som redegjort for tidligere i diskusjonen beskriver informant 1 EUs retningslinjer som lite detaljert og høyprofilert, noe som kan føre til en antakelse om at retningslinjene er utviklet for å være «brede» nok til at man kan si man har fulgt de, uten at de kan oversettes til en praksis som utfordrer oss. At retningslinjene er «brede», kan ifølge Larsson (2020) komme fra at under utvikling av retningslinjene, ble forbud mot visse bruksområder tonet ned, spesielt med påvirkning fra representantene fra industri, men også fordi juridiske spørsmål eksplisitt ble utelatt fra retningslinjene. Både informant 1 og Larsson (2020) deler synet om hvordan kommersielle interesser går framfor de etiske vurderingene i retningslinjene.

Fra informant 2 sitt tekniske perspektiv, er hun mer splittet på spørsmål om EU sin påvirkning, og kommersielle aktører sin påvirkning i utformingen av EUs retningslinjer. De siste årene har teknologiverden ifølge informant 2 vært dominert av techgiganter og plattformselskaper som ikke forholder seg til demokratiske spilleregler og gjør som de vil. Informant 2 mener at EU er positivt, fordi de har en aktiv teknologipolitikk. Hun beskriver detalj- og tilgangsnivået som passende, fordi at retningslinjer ikke kan være for teknologispesifikk, i et felt som fort blir utdatert. Informant 2 forteller at det er viktig å lage rom for store aktører når vi utarbeider retningslinjer for bruk av KI, fordi teknologisiden må få lov til å finne et marked, levere produkter, videreutvikle og tjene penger. Likevel understreker hun at det må være en balanse. Informant 2 belyser viktige nyanser i debatten om retningslinjenes funksjon. Samtidig, sett i lys av Larssons (2020) kritikk om at representanter fra industrien deltok i utformingen av retningslinjene, kan dette være en forklaring på hvorfor detalj- og tilgangsnivået i retningslinjene er begrenset. Selv om informant 2 argumenterer for prinsippene som passende, argumenter Larsson (2020) og informant 1, for at de er for generelle. Det er utfordrende å utvikle retningslinjer som alle aktører er fornøyd med, men om vi skal prioritere å ivareta det etiske vil ikke EUs KI prinsipper være tilstrekkelig, fordi prinsippene i seg selv feiler i å sikre de etiske dimensjonene når kommersielle interesser kommer framfor etiske vurderinger. De etiske verdiene må komme først, for at teknologien vi utvikler, bruker og implementerer skal gagne hele samfunnet.

Når vi gir industrirepresentanter tilgang til regulatoriske prosesser, skaper det en interessekonflikt mellom kommersielle målsetninger og etiske prinsipper som kan føre til at profitt går foran samfunnsmessige behov. Informant 1 sin påstand om at det er «for svakt» at Norges KI-etiske tilnærming er hentet fra EU, får støtte nettopp på grunn av at den tette koblingen til industrielle interesser kan komme i veien for samfunnets interesser. Basert på informant 1 sin beskrivelse av retningslinjene, kan dette få konsekvenser for implementeringen av KI i Norge. Hvis retningslinjene er så «brede» at kommersielle aktører kan hevde å ha fulgt dem uten reell etisk forankring, kan det samme gjelde for offentlige aktører, som beskrevet tidligere i diskusjonen. Selv om offentlige aktører ikke har en intensjon om å «etikkvaske» seg, kan de feilaktig anta at løsningene de utvikler og bruker er etisk forsvarlige, nettopp fordi de har fulgt de «brede» prinsippene. Informant 1 sitt argument understreker et behov for en nasjonalt tilpasset KI-etikk, som ikke kun følger EUs prinsipper,

men som også ivaretar norske verdier. Riegler, Lepperød og Røstad (2023) peker på et behov for bedre etisk kontroll og tilsyn av KI. Forfatterne skriver at store internasjonale selskaper i dag tar lite hensyn til data og etikk, fordi deres prioriteringer er styrt av kommersielle interesser, som fører til etisk tvilsomme utfall når det ikke utføres tilstrekkelig etisk tilsyn, som også beskrevet av Chun og Elkins (2023). Empirien framhever behovet for bedre etisk tilsyn i Norge, men reiser også spørsmål til hvordan vi skal regulere import av datasystemer.

En annen viktig debatt innen internasjonal påvirkning er knyttet til hvordan datasystemer ofte importeres. Begge informantene viste skepsis til kommersielle verktøy og import av KI-modeller. Informant 1 mener at om vi knytter dette til bias, vil det utgjøre en utfordring fordi importerte datasystem i noen tilfeller er utviklet av en viss gruppe privilegerte mennesker som ikke reflekterer mangfoldet vi har og ønsker i det norske samfunnet. Dersom KI-modeller som framover skal bli brukt i offentlig sektor er importerte, kan det bli utfordrende å evaluere om teknologien ivaretar norske verdier og etiske standarder, samt utfordringer knyttet til det å føre tilsyn på verktøyene, når vi ikke har nok innsikt. Informant 2 er også skeptisk til kommersielle verktøy, fordi vi har lite kontroll over dem når bedriftene bak ikke publiserer kodene. Derfor mener informant 2 at det er bedre å bygge egne analysemodeller, fordi da vet man hva modellene gjør, og har mye mer kontroll.

Selv om denne oppgaven ser på norsk offentlig sektor, er det også viktig å gi et innblikk i det større teknologiske bildet. Dette er viktig å få fram fordi norske borgere daglig bruker teknologi utviklet av store internasjonale selskap, som er «etisk tvilsom». Bergsjø og Strümke (2022), skriver at selskaper som Google og Facebook begge har praksiser som strider mot prinsippet om pålitelig og etisk forsvarlig KI. Selv om kommersielle aktører sin KI-etiske praktisering ofte er «tvilsom», mener informant 2 at om man skal bruke dem eller ikke, er et valg hver enkelt må ta. Hun forklarer at mange er villige til å ofre privatliv for økt bekvemmelighet, som å samtykke til posisjonstjenester på mobilen. For å bruke teknologi på en klok måte mener informant 2 at man trenger det hun omtaler som "ethical-AI-literacy" – en forståelse som omfatter de etiske aspektene ved kunnskapen som kreves for å bruke teknologi klokt. Selv om det er krevende å regulere kommersielle verktøy fra utlandet i Norge, kan økt digital kompetanse bidra til å minske dette gapet. På denne måten, selv om hver enkelt bruker står fritt til å velge hvordan de benytter kommersielle verktøy, kan høyere digital kompetanse

styrke evnen til å bruke teknologien på en mer bevisst og ansvarlig måte. Dette er et argument jeg vil komme tilbake til i 6.7.

6.4 Etisk tilsyn og ivaretakelse av KI-etiske prinsipper

Munn (2022) argumenterer for at KI-etiske prinsipper setter «normative idealer», og at man derfor kan stille seg skeptisk til deres funksjon, når de er vanskelig å oversette til praksis, et argument som støttes av Hagendorff (2020) som skriver at KI-etikken «svikter», som kan skyldes mangel på konsekvenser for å bryte retningslinjer. Informant 2 er mer splittet og mener derimot at prinsippene har en funksjon – fordi når vi har blitt enig om noe, så må politikken forholde seg til det. Hun påpeker at KI-etiske prinsipper gir en retning, og om noen går imot disse så kan man påtale det. Når vi har etablert at vi skal følge KI-etiske prinsipper, så må politikken også legge til rette for at man sikrer at disse overholdes. I nasjonal strategi for KI (kommunal- og moderniseringsdepartementet 2020) skriver regjeringen at «tilsynsmyndigheter skal føre kontroll med at systemer basert på kunstig intelligens på sitt tilsynsområde opererer innenfor prinsippene for ansvarlig og pålitelig bruk av kunstig intelligens». I dag har det ingen alvorlige konsekvenser om en norsk bedrift velger å ikke følge etiske retningslinjer og prinsipper, som stemmer overens med Munns (2022) analyse av hvordan rammeverk og prinsipper mangler midler for overholdelse, et argument som støttes av informant 1 som forklarer at oppfølging av etisk bruk av KI har vært «veldig» dårlig. Hun mener det er lite konsekvenser knyttet til det etiske utover omdømmet, fordi det er lovlig. Her finner informant 1 støtte hos Bergsjø og Strümke (2022), som mener vi må skape bedre strukturer for kontroll av systemer basert på KI, og hos Riegler, Lepperød og Røstad (2023) som mener vi ikke bør vente på EU sin KI-forordning, men at vi allerede nå bør etablere veilednings- og tilsynsorgan for KI i Norge.

Som Tasioulas (2022) påpeker, er spørsmål om regulering og håndhevelse grunnleggende etiske spørsmål, fordi de påvirker personlig frihet og rettferdighet og impliserer valg som reflekterer etiske verdier og prioriteringer. Dette perspektivet fremhever den komplekse naturen ved regulering av kunstig intelligens. Gjennom det empiriske grunnlaget ser man at dagens tilsyn av teknologi og ivaretakelse av KI-etiske prinsipper er mangelfullt. Videre spørsmål bør fokusere på hvordan teknologisk utvikling kan reguleres og håndheves på en

måte som både prioriterer det kollektive beste og samtidig kan overføres til praksis. Informant 1 etterspør et renere fokus på det etiske, fordi hun mener det etiske fort faller bort i diskusjoner om å regulere eller å ikke regulere, hvor etikken fort går over i det juridiske. På bakgrunn av argumentasjonen over, kunne det vært hensiktsmessig for Norge å utvikle mer detaljerte retningslinjer, slik at etiske verdier kan omsettes til verdifulle tiltak. Eller utvikle veiviser som fokuserer på etikk, som informant 1 etterlyser, som virksomheter som ønsker å ta i bruk KI kan bruke, altså bedre «prosessverktøy», som informant hun kaller det.

En gjentakende problemstilling i det empiriske grunnlaget er hvordan man skal delegere ansvar når noe går feil når vi bruker KI-verktøy. Som informant 2 sier, så har designvalg etiske konsekvenser. Hun mener at etikk handler om valg, og at valgene man tar har konsekvenser. Om de er klar over det eller ikke, sitter menneskene som utvikler teknologi på en stor makt. Informant 2 beskriver prioriteringer i designprosesser som «viktige», fordi prioritering av ett element kan gå på bekostning av de etiske vurderingene. Derfor mener informant 2 at mennesker innen tekniske fag og de som utvikler KI-modeller er en faktor av «moralske agenter», som hun påpeker det er for lav bevissthet om. Ifølge Hagendorff (2020, 115) burde teknologer ha høyere kunnskap om samfunnsteknologiske konsekvenser, for å få en høyere følelse av ansvarlighet for de moralske konsekvensene av arbeidet sitt. Men som informant 2 understreker er det ikke teknologene som tar alle valgene, og sier at organisasjonene som skal ta i bruk teknologien også bør være ansvarlig og evaluere teknologien. Å delegere ansvar når noe går feil, er en stor utfordring. Skal man plassere ansvar på systemutviklerne, brukere eller myndigheter? Dette er en utfordring Lanestedt, Goodwin og Andersen (2023) beskriver som et «av de vanskeligste temaene å komme til bunns i», men som samtidig krever tydelig avklaring for å opprettholde rettsikkerheten i det norske samfunn.

6.5 Etisk KI gjennom politisk vilje

Et viktig poeng i oppgavens diskusjon ble belyst i intervjuet med informant 2, som har en teknisk bakgrunn. Hun peker på at det er utfordrende å måle hvorvidt en teknologi faktisk oppfyller etiske krav. Hun mener det er lett å liste opp KI-etiske prinsipper og si at en løsning skal være *rettferdig*, men peker på det som krevende å finne ut om løsningen faktisk er det. En

løsning informant 2 belyser er at man kan rigge en studie som strekker seg over flere år, men at dette er en prosess som krever finansiering, samtykke og samarbeidspartnere, noe som gjør det «tungt, dyrt og krevende» å måle og evaluere teknologien for å senere ta etiske vurderinger. Informant 1 er enig i at etikken ikke gjør ting hverken «lettere, billigere, eller raskere», der vilje om å prioritere det etiske vil koste. Likevel mener hun at det er en omkostning virksomheter må være villige til å ta. Om det etiske skal prioriteres, mener informant 1 at det er noe det offentlige burde bestemme – altså at politikere skal si at vi skal velger den etiske løsningen, framfor den billigste og raskeste. Informant 1 påpeker at vi har oljemilliarder til å vurdere det etiske, så lenge det er politisk vilje. Regjeringen har vist politisk vilje ved å bevilge én milliard kroner til KI-forskning (Kunnskapsdepartementet 2024), men det gjenstår å se om disse midlene vil bli brukt til å prioritere de etiske aspektene som informant 1 etterlyser. Forskning tar tid – har vi tid til å vente når teknologiutviklingen skjer i et stadig raskere tempo? Som informant 1 peker på, vil det kreve politisk vilje å prioritere etiske hensyn. Hun bemerker at dagens politiske programmer inneholder få konkrete føringer for digitalisering og uttrykker at partiet Rødt burde formulert en annen tilnærming enn Høyre på dette området. Informant 1 etterlyser en politikk som tydeliggjør hvordan ulike partier forholder seg til digitalisering. Hun mener en mulig årsak til den lave prioriteringen av digitale spørsmål er den begrensede etterspørselen fra velgerne, noe hun knytter til mangel på kompetanse – et tema jeg vil utdype senere i diskusjonen.

Som informant 2 forteller at kommunene er presset til å ta i bruk teknologien uten ressurser til å sette seg inn i den. Hun uttrykker samme bekymring for små bedrifter, som ofte mangler både ressurser og forskningskapasitet til å sette seg inn i etiske prinsipper for kunstig intelligens, noe jeg tidligere har belyst i diskusjonen. KI-milliarden utlyst av regjeringen skal gå til forskningssentre, men ut ifra informant 2 sine observasjoner, er det også behov for mer ressurser i det offentlige, slik som for eksempel kommuner som skal implementere KI-teknologi.

Å prioritere etiske hensyn er ikke bare kostbart - det kan også innebære at allerede etablerte utdanninger må endres. Informant 2 forteller at det i helseutdanningene har vært utfordrende å integrere digital kompetanse, og dersom kunnskap om kunstig intelligens (KI) skal innlemmes, må andre fagområder nedprioriteres. Ifølge NORA og Digitaliseringsdirektoratets

(Digitaliseringsdirektoratet u.å) sin oversikt over KI-verktøy, var 40% av verktøyene i databasen rettet mot helsesektoren. Siden helse er et prioritert område, er det derfor særlig viktig å sikre etiske vurderinger, da sektoren håndterer store mengder sensitiv data – noe som skaper et økt behov for teknologi som ivaretar personvern og etikk. Informant 1 understreker på sin side at etiske hensyn bør inngå i flere utdanninger, for alle yrkesgrupper som vil bruke teknologi i arbeidslivet. Som informant 1 forteller, så trenger vi politisk vilje for å prioritere det etiske, dette vil også kreve politisk vilje for å bestemme at KI skal få en større del i utdanninger. Som regjeringen påpeker i den nasjonale strategien for KI, bør høyere utdanningsinstitusjoner vurdere hvordan etikk og personvern kan integreres i fag som datavitenskap og informatikk (Kommunal- og moderniseringsdepartementet 2020). For å gå fra oppfordring til et faktisk krav vil det imidlertid kreve politisk vilje. Og som informant 1 understreker, bør etiske hensyn få plass i flere utdanninger, ikke bare de tekniske som regjeringen redegjør for.

Med politisk vilje kan etiske vurderinger av KI få en mer sentral del i implementeringen av KI. Regjeringen skriver i nasjonal strategi at det er viktig at data har god kvalitet og struktur, og at hver enkelt virksomhet skal ha oversikt over hva datagrunnlaget de bruker betyr, brukes til, hvilke prosesser de inngår i, og om det finnes rettslig grunnlag for å dele dataene (Kommunal- og moderniseringsdepartementet). Slike evalueringer vil kreve ressurser, noe som informant 2 påpeker det mangler både i det offentlige og i privat sektor. Etikk koster, og informant 1 mener det er en omkostning bedrifter må ta. Selv om etiske vurderinger ikke gjør prosesser «rimeligere, raskere eller enklere», er det viktig at etikken står sentralt, som igjen - vil kreve politisk vilje.

6.6 Menneskene bak systemene

Som jeg har vist gjennom oppgaven, er etisk utvikling, bruk og implementering av teknologi en kompleks utfordring. KI opererer i stor skala og påvirker en rekke områder der teknologien kan føre til skade dersom den fatter urettferdige beslutninger (Ferrer mfl. 2021). Offentlig sektor er et særlig viktig område å analysere, ikke bare på grunn av dens tilgang til store mengder sensitiv data, men også fordi den har et grunnleggende ansvar for å ivareta borgernes sikkerhet og rettigheter. Ferrer mfl. (2021) skriver at KI stadig tar over viktige beslutninger

innenfor områder som helseevaluering og saksbehandling. Dersom KI-modellens beslutninger er urettferdige eller feil, kan det få alvorlige konsekvenser for både enkeltmennesket og for samfunnet. Et eksempel på dette er når automatiserte systemer systematisk diskriminerer en gruppe mennesker gjennom automatiseringsprosessen (Ferrer mfl. 2021). Dette kan skyldes dårlig representasjon i dataene som brukes til å trene algoritmene (Buolamwini og Gebru 2018), som i noen tilfeller kan være en refleksjon av hvem som utvikler algoritmene og hvordan deres verdenssyn gjenspeiles i teknologien (D'Ignazio og Klein 2020; Borenstein og Howard 2017; Mullaney og Hicks 2021). Gjennom data og utviklingsprosesser kan eksisterende bias i samfunnet forsterkes gjennom automatisering (Ntoutsi mfl. 2020; Mehrabi mfl. 2021). D'Ignazio og Klein (2020) framhever at mennesker som utvikler KI-systemer ofte tilhører en dominant gruppe – noe som kan føre til at deres verdenssyn reflekteres i løsninger, og at andre perspektiver blir ekskludert, som informant 1 også beskriver som en bekymring. Om man ikke har mangfold i hvem som lager KI-systemer, vil sosiale, økonomiske og historiske skjevheter kunne videreføres og forsterkes gjennom automatisering (Ntoutsi mfl. 2020, Mehrabi mfl. 2021, D'Ignazio og Klein 2020).

For å sikre etiske hensyn i norsk KI-implementering vil det derfor mangfold være en viktig prioritering når vi utvikler nye KI-modeller. Dette kan også knyttes opp mot den tidligere diskusjonen om hvordan Norge importerer løsninger. Selv om vi i Norge har mulighet til å påvirke teknologi som vi selv utvikler, har vi lite kontroll over hvem som utvikler verktøyene vi importerer. Informant 1 forteller at flere KI-modeller vi bruker i Norge, er importert fra land som har mindre likestilling enn Norge, noe hun mener utfordrer de humanistiske grunnverdiene som menneskeverd, likeverd, kritisk tenking og demokrati. Dette er en sentral bekymring som regjeringen ikke har tilstrekkelig redegjort for i nasjonal strategi for KI (Kommunal- og moderniseringsdepartementet 2020). Hvordan KI-modeller er utviklet har stor betydning for det etiske, derfor er det viktig å ikke kun fokusere på *hvordan* algoritmer utvikles, men også *hvem* som utvikler de. Det empiriske grunnlaget understreker et behov for mangfold i hvem som utvikler KI-systemer. Som litteraturen viser, er det viktig med mangfold av kjønn, alder og etnisitet (D'Ignazio og Klein, Buolamwini og Gebru 2018, Mehrabi mfl. 2021), men det vil også være viktig med mangfold av tanker knyttet til det faglige. Dette vil jeg bygge videre på i kapittel 6.8.

Avsnittene over redegjør for hvorfor det er viktig å fokusere på *hvem* som utvikler systemene, mens i de neste avsnittene vil jeg derimot dreie fokuset mot *hvordan* systemene utvikles. Jeg vil belyse dette fokuset ved å diskutere hvilken rolle data brukt for å trene KI-modeller, har i den KI-etiske debatten. Som Buolamwini og Gebru (2018) sin studie «Gender Shades» illustrerte, kan mangel på mangfold i datagrunnlag få alvorlige konsekvenser – og er et konkret eksempel på hvorfor det er viktig med mangfold i datainnsamlingsprosesser. Informant 2 med teknisk bakgrunn og erfaring med å analysere data og datasett, påpeker at datasett som blir brukt for å trene algoritmer, ikke alltid er representativt for befolkningen. Hun forteller at man vet mer om noen grupper enn om andre, noe som kan føre til at man utvikler løsninger for det området og de gruppene man kjenner til. Dette understreker at mangfold i datagrunnlag som illustrert av Buolamwini og Gebru (2018), også er viktig i norsk kontekst. Som beskrevet tidligere, er dette en prioritering som vil kreve ressurser, men en prioritering vi må akseptere om vi ønsker etisk gode verktøy.

6.7 Digital kompetanse

Økt digital kompetanse har vært et gjennomgående tema i denne oppgaven (Chun og Elkins 2023, Lanestedt, Goodwin og Andersen 2023, Kommunal- moderniseringsdepartementet 2020, Bergsjø og Strümke 2022, Riegler, Lepperød og Røstad 2023). Som jeg har beskrevet i oppgaven og i diskusjonen, er bias en sentral bekymring innenfor KI-etikken. Informant 2 jobber mye med å avdekke bias i sitt arbeid der hun analyserer data og datasett brukt for å trene KI-modeller. En bias hun har observert, er knyttet til hvordan innbyggere klarer å bruke KI-verktøy. Offentlige tjenester skal være universelle og tilgjengelige for alle, i motsetning til de kommersielle. Selv om dette argumentet illustrerer hvorfor det er viktig å tilrettelegge tjenester for brukere, kan det også ses i sammenheng med et behov for økt digital kompetanse. Som jeg har beskrevet tidligere i diskusjonen, er åpenhet et viktig norsk kjerneprinsipp som regjeringen skriver i nasjonal strategi - som sier at borgere skal ha rett på å vite hvordan beslutninger som påvirker dem fattes (Kommunal- og moderniserinsdepartementet 2020). Selv om en viktig faktor for å sikre åpenhet vil være at offentlig sektor gir innsyn, kan man også oppnå høyere grad av åpenhet om det digitale kompetansenivået i Norge heves. KI og teknologi er komplisert, og for å kunne forstå hvorfor KI-modeller fatter de beslutningene de gjør, kan økt forståelse av algoritmiske prosesser bidra i å tette dette gapet. Algoritmer blir av

flere beskrevet som en «sort boks» med lav forklaringsevne (Kommunal- og moderniseringsdepartementet 2020). Selv om det er vanskelig å gi innsikt i hvordan algoritmene fungerer i praksis, kan innføring i sentrale teknologiske prosesser som for eksempel datainnsamling, gi en økt forståelse for brukere (kommunal- og moderniseringsdepartementet 2020).

I begge intervjuene etterspurte informantene mer digital kompetanse. Informant 1 etterlyser spesifikt mer kompetanse som ser på de etiske aspektene om KI, som hun mener er viktig for å kunne bruke teknologien klokt, og ønsker at de som går ut av skolen framover skal være godt rustet i denne tematikken. Riegler, Lepperød og Røstad (2023) mener vi må øke kunnskapen i samfunnet ved å innføre opplæring i ansvarlig bruk av KI i alle utdanninger. Det vil også være viktig at menneskene som skal ta i bruk verktøyene har høy nok digital kompetanse, for å i større grad kunne ekskludere farer for tolkningsbias (Danks og London 2017, Ferrer mfl. 2021). Datatilsynets sluttrapport av NAV sitt prøveprosjekt (Datatilsynet 2022), diskuterer hvordan beslutningsstøtte skal bli brukt på en god og riktig måte. For at veiledere skal kunne vurdere modellens beslutninger på et selvstendig og trygt grunnlag, ser man et behov for at veiledere burde få opplæring og instrukser i hvordan algoritmen fungerer og brukes. Dette mener Datatilsynet kan bidra til at veiledere kan identifisere diskriminering, forskjellsbehandling og feil (Datatilsynet 2022). Som informant 1 sier, så trenger vi etisk kompetanse for å bruke teknologien klokt, og Datatilsynets rapport viser hvordan dette kan se ut i praksis, og hvilke fordeler økt digital kompetanse kan gi.

Økt digital kompetanse blant forskere vil også være viktig. Som Chun og Elkins (2023) påpeker, er digital kompetanse nødvendig for å kunne bidra konstruktivt til store etiske spørsmål. Informant 1 forteller at teknologer og humanister ofte ser ulike utfordringer. Med høyere digital kompetanse hos humanister og økt etisk forståelse blant teknologer, kan evalueringen av etiske dimensjoner bli enklere. Dette kan oppnås gjennom tverrfaglig samarbeid, som jeg vil diskutere i neste delkapittel.

6.8 Tverrfaglig samarbeid

Behov for tverrfaglig samarbeid, har gjentatt seg gjennom hele oppgaven, både i intervjuene og i litteraturundersøkelsen (Frank et al 2019, Chun og Elkins 2023, Dimock 2020, Riegler, Lepperød og Røstad 2023, Kazim og Kosyiyama 2021, Ferrer et al 2021). Dette reflekterer et tydelig behov for å gå inn i de tverrfaglige debattene for å sikre etisk KI.

Informant 1 beskriver at humanistiske verdier er noe hun ser mindre av i de tekniske miljøene, der perspektiver som likeverd og likestilling ikke tematiseres. Informant 1 ønsker at flere går inn i de tverrfaglige samtalene, og mener at teknologene med sin kompetanse kan oppdage etiske problemer humanister ikke kan se. Informant 2 ser også et behov for tverrfaglig samarbeid, og ønsker at teknologene skal lære mer om etiske perspektiver, og humanistene mer om tekniske perspektiv. Informant 2 mener at: «Ideelt sett så burde humanister samfunnsvitere og alle inn så tidlig som mulig i utviklingen tenker jeg. Ikke bare når det er ferdig å slippes ut». Hun mener at de tverrfaglige samtalene må skje tidlig for å få god effekt, fordi identifisering av problemer sett fra utsiden har lite verdi for teknologene etter at prosessen er ferdig. I pressemeldingen fra Kunnskapsdepartementet om KI-milliarden står det at tverrfaglig forskning vil bli prioritert (Kunnskapsdepartementet 2024). Ettersom min forskning har vist et sterkt behov for de tverrfaglige debattene, blir det spennende å se hvilke resultater denne prioriteringen vil føre til.

For å sikre at KI skal bidra til å tjene «menneskehetens kollektive behov», skriver Dimock (2020) at menneskene bak systemene må representere menneskeheten – noe som krever mangfold av tanker, etnisitet, kjønn, kultur, alder, nasjonalitet og disipliner. Etiske spørsmål bør vurderes av et mangfold av perspektiver, som Chun og Elkins (2023) skriver. Om KI integreres i hele universitetsstrukturen, mener Chun og Elkins at det kan bidra til å viske ut grensen mellom teknologi og humaniora. Min forskning viser at de tverrfaglige diskusjonene bør foregå gjennom hele prosessen: utvikling, forskning, bruk og implementering. Dette vil være viktig for å sikre at KI brukes, utvikles og implementeres på en etisk forsvarlig måte.

Kapittel 7: Konklusjon

Denne masteroppgaven har hatt som mål å besvare følgende problemstilling:

> *Hvordan kan vi sikre at etiske hensyn står sentralt i implementeringen av kunstig intelligens i norsk offentlig sektor?*

Oppgaven har identifisert en rekke utfordringer som utfordrer et sentralt fokus på etiske hensyn i implementeringen av KI i offentlig sektor. Den mest sentrale bekymringen i denne oppgaven er knyttet til KI-etiske prinsipper. De er abstrakte, tvetydige, de mangler sosiale og politiske kontekster, og ivaretakelse av ett etisk prinsipp kan gå på bekostning av ivaretakelsen av et annet, noe som skaper et gap mellom teori og praksis. Dette fører til variasjon i etisk praksis på tvers av sektorer, og gjør det utfordrende å omsette prinsippene til verdifulle tiltak. Som informantene og litteraturundersøkelsen har demonstrert, sikrer ikke dagens nasjonale KI-etiske tilnærming etiske hensyn. Den legger føringer for at vi ønsker oss etiske løsninger, men den klarer ikke å realisere dette målet.

Gjennom en omfattende litteraturundersøkelse og dybdeintervjuer, viser min forskning fire måter for å kunne bevare «det etiske» i de etiske diskusjonene, og å sikre at KI brukes på en rettferdig måte som gagner hele samfunnet. Disse fire måtene er at vi må øke digital kompetanse, fremme tverrfaglige samarbeid, konkretisere KI-etiske prinsipper, og sikre at prinsippene gjenspeiler norske verdier.

Det første punktet er å øke den digitale kompetansen i det norske samfunnet. Det er et behov for både KI-etisk kompetanse, men også generell KI-kompetanse for hvordan teknologien fungerer i praksis. Det vil være vanskelig å stille kritiske spørsmål til teknologien, om man ikke har en grunnforståelse av teknologiske prosesser – som at algoritmer er trent på store datasett som kan reflekteres i algoritmens utfall. Det er behov for KI-rettet kompetanse i utdanninger i og utenfor teknologi, men man må også rette fokus mot mennesker som allerede er ferdig utdannet og i arbeid. Nå som teknologien stadig tar en større rolle innenfor offentlig sektor, vil det bli viktig at menneskene som skal ta i bruk KI-modeller som beslutningsstøtteverktøy, opparbeider seg høyere digital kompetanse. Dette vil gjøre arbeidere

bedre rustet til kritisk tenkning og til å forstå teknologiens begrensninger for å kunne foreta etiske vurderinger.

Det andre punktet er tverrfaglig samarbeid, både innen forskning og i selve utviklingen av KI-modeller. Humanister og samfunnsvitere (og andre fagfelt) har en annen kompetanse enn teknologene. De kan evaluere KI-systemenes påvirkning på mennesker og samfunn fra andre perspektiver enn det tekniske. Det er viktig at dette samarbeidet skjer tidlig i prosesser, for å få størst effekt. Det vil være mer effektivt å evaluere teknologien i en tidlig fase. Det vil også være mer økonomisk i lengden, fordi etisk evaluering i etterkant er dyrt og tidkrevende. Ideelt sett er alle parter med både før, i og etter prosessen, slik at det vil bli mer håndterbart å utvikle etiske løsninger. Det vil også bidra til å bevare det etiske – fordi gjennom samarbeid deler og lærer man av hverandre. Humanistene vil få en bedre forståelse av tekniske prosesser, og teknologene vil få høyere kunnskap i etiske perspektiver.

Det tredje punktet for å kunne bevare det etiske, er at vi trenger konkretisering av etiske prinsipper og retningslinjer, for å sikre lik etisk praksis på tvers av sektorer. Her peker min forskning på flere mulige retninger man kan gå i: man kan utvikle egne retningslinjer, man kan bygge på EU sine prinsipper til å passe den norske samfunnsmodellen og ivareta norske verdier, samt at man kan legge frem konkrete og tydelige veivisere for bruk av KI og bruk av dens prinsipper og retningslinjer. Min forskning viser problemet vi står ovenfor, og at tiltak må settes inn, men er utydelig i nøyaktig hvordan dette skal foregå. Dette kan forstås som et tegn på at etisk KI ikke er så enkelt å operasjonalisere i praksis.

Det fjerde punktet er knyttet til den store spenningen i etisk implementering av KI. Vi er enige i at teknologien skal bygge på etiske prinsipper og verdier, men likevel er det stor uenighet om hvordan dette ser ut i praksis. Dette gjør det utfordrende å sikre at vi ivaretar det etiske. Det vil være viktig å legge inn en stor satsing på området for å kunne drøfte styrker, svakheter, muligheter og trusler, og hvordan man på best mulig måte kan identifisere hva det vil si at kunstig intelligens er etisk. I 2023 lanserte regjeringen KI-milliarden, en finansieringsordning som skal støtte forskningssentre dedikert til kunstig intelligens. Målet er å fremme tverrfaglig forskning på sentrale problemstillinger innen KI-feltet. Selv om KI-milliarden er et viktig økonomisk tiltak, har denne oppgaven vist et behov for ressurser også

på andre områder. For at mennesker som utvikler, bruker og implementerer teknologien skal kunne handle i tråd med etiske verdier, må det legges til rette for at de har tilstrekkelige ressurser til å gjøre det, i form av kompetanse, utdanning, erfaring og tid. Etikk gjør ikke prosesser «enklere, billigere eller raskere», men det er en kostnad vi må akseptere for å sikre at etiske hensyn står sentralt ved implementeringen av kunstig intelligens i offentlig sektor.

7.1 Begrensninger og videre forskning

Denne oppgaven analyserer teknologi gjennom en kritisk linse, noe som kan overskygge positive aspekter ved KI-implementering i norsk offentlig sektor. Datagrunnlaget og oppgavens teori inkluderer både humanistiske og tekniske perspektiver, men med et overordnet humanistisk fokus, kan en konsekvens være en mindre nyansert debatt. Dette understreker et av oppgavens hovedpoeng, nemlig behovet for tverrfaglig samarbeid. Som humanist mangler jeg teknologisk kompetanse til å evaluere teknologien fra et teknisk perspektiv. Uten teknologisk kompetanse, er det vanskelig å evaluere selve teknologien for å videre kunne evaluere etiske problemstillinger. Mitt hovedfokus har derfor vært å fremheve sentrale etiske utfordringer ved KI og deres innvirkning på den KI-etiske debatten i norsk kontekst. Videre må det nevnes at denne oppgaven har to intervjuer, noe som begrenser generaliserbarheten, fordi forskere innenfor de teknologiske og humanistiske feltene har ulike meninger. Likevel mener jeg at funn gjort i litteraturundersøkelsen støtter opp mot funnene i intervjuene, hvor intervjuene bidro til å knytte funnene opp mot norsk implementering av KI i offentlig sektor.

Videre forskning bør rette seg mot den tverrfaglige debatten, der samarbeid med forskere fra ulike fagfelt kan bidra til å utvikle gode tilnærminger til KI-etiske bekymringer, for den norske konteksten. Gjennom en tverrfaglig tilnærming kan spørsmål denne oppgaven stiller, besvares. Dette vil sikre at etiske perspektiver får en større rolle innenfor områder der fokuset tidligere har vært på tekniske aspekter. En tverrfaglig tilnærming til evalueringen av KI-etiske prinsipper vil være avgjørende. Det har vokst fram en global og nasjonal enighet om at teknologiske løsninger skal være både gode og etiske. Men hva innebærer det at en løsning er etisk? Internasjonalt, men også i Norge, har utviklingen og bruken av «myke» styringsmekanismer, som KI-etiske prinsipper, blitt en strategi for å sikre etisk bruk av KI i

offentlig sektor og industri. Som denne oppgaven har vist, er ikke disse prinsippene tilstrekkelig for å møte dette ønsket. Dersom Norge skal fortsette med en prinsippbasert tilnærming, er det behov for mer forskning på prinsippenes virkning, samt en grundig vurdering av hvordan de kan bidra positivt. Det er også nødvendig med videre forskning på hvordan gapet mellom teori og praksis kan lukkes, og hvordan norske verdier kan reflekteres i prinsippene. For å kunne svare på alle disse spørsmålene, vil det kreve grundig forskning, på hvordan KI-etiske prinsipper og retningslinjer oppfattes i brukeres og utvikleres praktiske hverdag. Og, kartlegging på hvilke utfordringer de som skal implementere teknologien opplever i møte med prinsippene. Her vil det være spesielt viktig å ha et høyt fokus på de aktørene som i utgangspunktet har lite ressurser til å sette seg inn i prinsippene, fordi vi trenger gode etiske løsninger på alle nivå.

Et annet punkt som trenger mer forskning, er det økende behovet for digital kompetanse. Den første ideen til dette forskningsprosjektet, var å se nærmere på kompetansebehovet for de som skal ta i bruk teknologien. Vi trenger forskning både for å evaluere kompetansebehovet, og for å identifisere hva denne kompetansen skal innebære. Denne ideen viste seg dessverre å være for omfattende for dette forskningsprosjektet, men det er noe jeg håper å kunne forske videre på i fremtiden. En tilnærming til denne problemstillingen, er å starte med å kartlegge det eksisterende nivået av digital kompetanse og videre kartlegge utfordringene og bekymringene blant de som skal ta i bruk KI-verktøy. Dette vil kreve omfattende forskning, inkludert samarbeid med flere offentlige aktører og ansatte på ulike nivåer i virksomhetene. Neste steg vil være å undersøke hvilket kompetansenivå som er nødvendig for at brukerne skal kunne anvende teknologien på en trygg og effektiv måte, samt utvikle strategier for hvordan denne kompetansen best kan formidles. Teknologisk utvikling påvirker mange aspekter av samfunnet, arbeidsliv, utdanning, demokrati og sosial deltakelse. Økt digital kompetanse kan gjøre befolkningen bedre rustet til å bruke teknologien på en effektiv, trygg og kritisk måte. Denne masteroppgaven er spesifikt rettet mot offentlig sektor, men økt digital kompetanse vil også være viktig for borgeres private liv.

8. Litteraturliste

Anker, Trine. 2021. *Analyse i praksis: En håndbok for masterstudenter*. Oslo: Cappelen Damm Akademisk.

Baste, Øystein Flø, Alexandra Schultz, og Jens Andersen Osberg. 2023. «Mens vi venter på at EU skal regulere kunstig intelligens.» *Stat & Styring*, 33 (3): 15-20. https://www.idunn.no/doi/10.18261/stat.33.3.3

Bergsjø, Leonora, og Inga Strümke. 2022. «Etter to år med nasjonal strategi for kunstig intelligens trengs opplæring og struktur.» *Digi.no*, 15. Januar, 2022. https://www.digi.no/artikler/debatt-etter-to-ar-med-nasjonal-strategi-for-kunstig-intelligens-trengs-opplaering-og-struktur/516538

Berry, David. 2022. «AI, ethics, and digital humanities.» I *The Bloomsbury Handbook to the Digital Humanities*, redigert av James O'Sullivan, 445-557. London: Bloomsbury Handbooks.

Bromfield, Heather og Mona Naomi Lintvedt. 2022. «Snubler Norge inn i en algoritmisk velferdsdystopi?» *Tidsskrift for velverdsforskning*. 25 (3): 1–15 https://doi.org/10.18261/tfv.25.3.2

Buolamwini, Joy, og Timnit Gebru. 2018. «Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.» *Conference on Fairness, Accountability and Transparency*, 81: 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html?mod=article_inline.

Chun, Jon, og Kathrine Elkins. 2023. «The Crisis of Artificial Intelligence: A New Digital Humanities Curriculum for Human-Centered AI.» *International Journal of Humanities and Arts Computing* 17 (2): 147–167. https://www.euppublishing.com/doi/10.3366/ijhac.2023.0310

Coeckelberrgh, Mark. 2020. *AI Ethics*. Cambridge: The MIT Press.

Cormen, H. Thomas, Charles E. Leiserson, Ronald L. Rivers og Clifford Stein. 2022. *Introduction to Algorithms.* 3. utgave. Cambridge: The MIT Press.

Dalland, Olav. 2020. *Metode og oppgaveskriving*. 7. utg. Oslo: Gyldendal.

Datatilsynet. 2020. «Rammeverk for Datatilsynets regulatoriske sandkasse for kunstig intelligens.». Hentet 12. September 2024. https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/rammeverk-for-den-regulatoriske-sandkassen/

Datatilsynet. 2022. «NAV - sluttrapport.» Hentet 27. August 2024. https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/ferdige-prosjekter-og-rapporter/nav-sluttrapport/.

Datatilsynet. 2023. «Om personopplysningsloven med forordning og når den gjelder.» Hentet 25. August 2024. https://www.datatilsynet.no/regelverk-og-verktoy/lover-og-regler/om-personopplysningsloven-og-nar-den-gjelder/.

Datatilsynet. 2024. «Etikk: Rammer for ansvarlig KI.» Hentet 20. August 2024. https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/ferdige-prosjekter-og-rapporter/politihogskolen-sluttrapport-prevbot/etikk-rammer-for-ansvarlig-ki/.

Datatilsynet. u.å. «Sandkassesiden.» Hentet 20. juli 2024. https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/.

Danks, David, og Alex J. London. 2017. «Algorithmic Bias in Autonomous Systems.» *International Joint Conference on Artificial Intelligence* 26: 4691–4697. https://www.cmu.edu/dietrich/philosophy/docs/london/IJCAI17-AlgorithmicBias-Distrib.pdf

Delua, Julianna. 2021. «Supervised Versus Unsupervised Learning: What's the Difference?.» *IBM*. https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning.

Digitaliseringsdirektoratet. u.å. «Bruk av Kunstig Intelligens i offentlig sektor.» Hentet 01. Mai 2024. https://www.digdir.no/rikets-digitale-tilstand/bruk-av-kunstig-intelligens-i-offentlig-sektor/4463.

D'Ignazio, Catherine, og Lauren F. Klein. 2020. *Data Feminism*. Westmont: MIT Press.

Dimock, W. C. 2020. «AI and the Humanities.» *PMLA* 135 (3): 449–454. https://doi.org/10.1632/pmla.2020.135.3.449.

EITCA. 2023. «Hva er estimatorene?» *Europeisk IT-sertifiseringsinstitutt*. Oppdatert 2. September, 2023. https://no.eitca.org/artificial-intelligence/eitc-ai-gcml-google-cloud-machine-learning/first-steps-in-machine-learning/plain-and-simple-estimators/what-are-the-estimators/.

Ferrer, Xavier., Tom van Nuenen, Jose M. Such, Mark Coté, og Natalia Criado. 2021. «Bias and Discrimination in AI: a Cross-disciplinary Perspective.» *IEEE Technology and Society Magazine* 40 (2): 72-80. https://doi.org/10.1109/MTS.2021.3056293

Frank, M. R., D. Wang, M. Cebrian, mlf. 2019. «The Evolution of Citation Graphs in Artificial Intelligence Research.» *Nature Machine Intelligence* 1: 79–85. https://doi.org/10.1038/s42256-019-0024-5.

Hagendorff, Thilo. 2020. «The Ethics of AI Ethics: An Evaluation of Guidelines.» *Minds and Machines* 30 (1): 99-120. https://doi.org/10.1007/s11023-020-09517-8.

Haalama, Jaana og Tania Kalliokoski. 2022. «AI Ethics as Applied Ethics.» *Frontiers in computer science* 4:776837. https://doi.org/10.3389/fcomp.2022.776837.

Hickok, M. 2021. «Lessons Learned from AI Ethics Principles for Future Actions.» *AI and Ethics* 1 (1): 41–47. https://doi.org/10.1007/s43681-020-00008-1.

HLEG. 2019. «Ethics Guidelines for Trustworthy AI.» *European Commission*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Howard, Ayanna., og Jason Borenstein. 2017. «The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity.» *Science and Engineering Ethics* 24: 1521–1536. https://doi.org/10.1007/s11948-017-9975-2.

IBM. u.å. «What is AI ethics? » Hentet 12. Juni 2024. https://www.ibm.com/topics/ai-ethics

Jobin, Anna, Marcello Ienca, og Effy Vayena. 2019. «The Global Landscape of AI Ethics Guidelines.» *Nature Machine Intelligence* 1 (9): 389–399. https://doi.org/10.1038/s42256-019-0088-2.

Kazim, Emre og Adriano Koshiyama. 2021. «A High-Level Overview of AI Ethics.» *Patterns* 2 (9): 100314. https://doi.org/10.1016/j.patter.2021.100314.

Kirkpatrick, Keith. 2016. «Battling Algorithmic Bias: How Do We Ensure Algorithms Treat Us Fairly?» *Communications of the ACM* 59 (10): 16–17. http://dx.doi.org/10.1145/2983270.

Kommunal- og moderniseringsdepartementet. 2020. «Nasjonal Strategi for Kunstig Intelligens.» *Regjeringen.no.* https://www.regjeringen.no/no/dokumenter/nasjonal-strategi-forkunstigintelligens/id2685594/?ch=1.

Kvale, Steinar, og Svend Brinkman. 2009. *Det kvalitative forskningsintervju*. 2. utgave. Oslo: Gyldendal akademisk.

Kvale, Steinar og Svend Brinkmann. 2015. *Det kvalitative forskningsintervju*. 3. utg. Oslo: Gyldendal akademisk

Kunnskapsdepartementet og Statsministerens kontor. 2023. «Regjeringen med milliardsatsing på kunstig intelligens». Regjeringen. Hentet 6. juli 2024. https://www.regjeringen.no/no/aktuelt/regjeringen-med-milliardsatsing-pa-kunstig-intelligens/id2993214/

Kunnskapsdepartementet. 2024. «No Kjem Utlysinga av KI-Milliarden.» Regjeringen. Hentet 4. juli 2024. https://www.regjeringen.no/no/aktuelt/no-kjem-utlysinga-av-ki-milliarden/id3030861/.

Kvale, Steinar, og Svend Brinkmann. 2015. *Det Kvalitative Forskningsintervju.* 3. utg. Oslo: Gyldendal Akademisk.

Lanestedt, Gjermund., Morten Goodwin, og Per-Arne Andersen. 2023. «Tid for en (Mer) Intelligent Statsforvaltning?» *Stat & Styring* 33 (3): 7–14. https://www.idunn.no/doi/10.18261/stat.33.3.2.

Larson, Jeff, Surya Mattu, Lauren Kirchner, og Julia Angwin. 2016. «How We Analyzed the COMPAS Recidivism Algorithm.» *ProPublica*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Larsson, Stefan. 2020. «On the Governance of Artificial Intelligence Through Ethics Guidelines.» *Asian Journal of Law and Society* 7 (3): 437–451. https://doi.org/10.1017/als.2020.19.

Leavy, Susan, Barry O'Sullivan, og Eugenia Siapera. 2020. «Data, Power and Bias in Artificial Intelligence.» *arXiv preprint.* https://doi.org/10.48550/arXiv.2008.07341.

Li, Fei-Fei, og John Etchemendy. «Welcome to the Stanford Institute for Human-Centered Artificial Intelligence.» *Stanford University*. https://hai.stanford.edu/navigate/welcome.

Madan, Rohit og Mona Ashok. 2023. «AI Adoption and Diffusion in Public Administration: A Systematic Literature Review and Future Research Agenda.» *Government Information Quarterly* 40 (1): 101774. https://doi.org/10.1016/j.giq.2022.101774.

Mahesh, Batta. 2020. «Machine Learning Algorithms - A Review.» *International Journal of Science and Research (IJSR)* 9 (1): 381–386. http://dx.doi.org/10.21275/ART20203995.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, og Aram Galstyan. 2021. «A Survey on Bias and Fairness in Machine Learning.» *ACM Computing Surveys* 54 (6): 1–35. https://doi.org/10.1145/3457607.

Morley, Jessica, Anat Elhalal, Francesca Garcia, Libby Kinsey, Jakob Mökander, og Luciano Floridi. 2021. «Ethics as a Service: A Pragmatic Operationalisation of AI Ethics.» *Minds and Machines* 31 (2): 239–256. https://doi.org/10.1007/s11023-021-09563-w.

Mullaney, Thomas. 2021. «Introduction.» I *Your Computer Is On Fire,* redigert av Thomas S. Mullaney, Benjamin Peters, Mar Hicks og Kavita Philip. Cambridge: The MIT Press

Muhaisen, Sahara. 2024. «Regjeringen ber 80 prosent av offentlig sektor bruke KI innen 2025.» *NRK.no*, 16. april, 2024. https://www.nrk.no/norge/regjeringen-vil-at-80-prosent-av-offentlig-sektor-bruker-ki_-_-urealistisk_-mener-ki-forsker-1.16843972.

Munn, Luke. 2023. «The Uselessness of AI Ethics.» *AI and Ethics* 3: 869-877 https://doi.org/10.1007/s43681-022-00209-w.

Ntoutsi, Eirini, Pavlos Fafalios, Ujwal Gadiraju, Vasielios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, og Salvatore Ruggieri. 2020. «Bias in Data-Driven Artificial Intelligence Systems—An Introductory Survey.» *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10 (3): e1356. https://doi.org/10.1002/widm.1356.

Rahwan, Iyad. 2018. «Society-in-the-Loop: Programming the Algorithmic Social Contract.» *Ethics and Information Technology* 20: 5–14. https://doi.org/10.1007/s10676-017-9430-8.

EØS-notatbasen. 2024. «Forslag til Forordning om Kunstig Intelligens (KI-Forordningen).» Regjeringen. https://www.regjeringen.no/no/sub/eos-notatbasen/notatene/2021/juni/forslag-til-forordning-om-kunstig-intelligens-ki-forordningen/id2884935/

Riegler, Michael A., Mikkel Lepperød, og Lillian Røstad. 2023. «Norge som fanebærer for etikk i en uforutsigbar framtid.» *Digi.no*, 24. Desember, 2023. https://www.digi.no/artikler/debatt-norge-som-fanebaerer-for-etikk-i-en-uforutsigbar-framtid/541427.

Riegler, Michael A, og Olav Lysne. 2024. «EU-Lov om kunstig intelligens godkjent – hva burde Norge gjøre nå?» *Digi.no*, 13. April, 2024. https://www.digi.no/artikler/debatt-eu-lov-om-kunstig-intelligens-godkjent-hva-burde-norge-gjore-na/545682.

Reutter, Lisa Marie, og Heather Broomfield. 2019. «Kunstig Intelligens/Data Science: en kartlegging av status, utfordringer og behov i norsk offentlig sektor - første resultater.» Forskningsrapport, *NTNU*. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2634733.

Siau, Keng, og Weiyu Wang. 2020. «Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI.» *Journal of Database Management* (JDM) 31 (2): 74–87. https://www.igi-global.com/article/artificial-intelligence-ai-ethics/249172.

Sikt. U.å. «Informasjon til Deltakarane i Forskningsprosjekt.» https://sikt.no/tjenester/personverntjenester-forskning/fylle-ut-meldeskjema-personopplysninger/informasjon-til-deltakarane-i-forskingsprosjekt.

Tasioulas, John. 2022. «Artificial Intelligence, Humanistic Ethics.» *Daedalus*, *151*(2), 232-243. https://doi.org/10.1162/daed_a_01912.

Universitetet i Bergen. 2024. «RETTE - UiB's Prosjektoversikt.» Hentet 1. oktober 2024. https://www.uib.no/forskningsetikk/128207/rette-uibs-prosjektoversikt#hvilke-tillatelser-trenger-jeg-for-behandle-personopplysninger-nbsp-nbsp.

Wang, Pei. 2019. «On Defining Artificial Intelligence.» *Journal of Artificial General Intelligence 10* (2): 1–37. https://doi.org/10.2478/jagi-2019-0002.

Whang, E. Steven, Yuji Roh, Hwanjun Song, og Jea-Gil Lee. 2023. «Data Collection and Quality Challenges in Deep Learning: A Data-Centric AI Perspective.» *The VLDB Journal* 32 (4): 791–813. https://doi.org/10.1007/s00778-022-00775-9.

Vedlegg

Vedlegg 1: Intervjuguide

1) Hva er ditt forhold til KI-etikk?
2) Det er mange ulike meninger om hva som går innenfor KI-etikk, og betydningen til etikk begrepet i sammenheng med KI, kan du fortelle litt om dette?
3) Per nå, hva ser du på som de største utfordringene innenfor KI-etikk?
4) Innenfor KI-etikk er det høyt fokus på bias/skjevhet, hva må vi være spesielt oppmerksomme på, for å unngå skjevhet når vi implementerer KI i norsk offentlig sektor.
5) Hvor viktig er det at folk som tar i bruk KI-teknologi er kjent med en algoritmisk prosess, for å kunne utelukke skjevhet?
6) Er dagens retningslinjer og regulering av KI i Norge gode nok?
7) Hva mener du vil være de viktigste skrittene mot en etisk implementering av teknologien?
8) I den nasjonale strategien for kunstig intelligens, kommer det tydelig fram at mye av den norske etiske tilnærmingen er hentet fra EU og HLEG sitt arbeid - hva tenker du om det?
9) I hvilken del av norsk offentlig sektor, vil det være etisk mest problematisk å ta i bruk KI-teknologi? og hvor kan teknologien være enklest?
10) Innenfor KI-etikk litteratur er det flere forskere som peker til hvordan retningslinjer og regulering av teknologien ikke er tilstrekkelig for å sikre at teknologien blir riktig brukt, har du noen tanker om andre tiltak eller tilnærminger som kan bidra i å sikre etisk bruk av KI?

Vedlegg 2: Samtykkeskjema

Forespørsel om deltakelse i forskingsprosjekt om KI

”Etisk implementering av kunstig intelligens i norsk offentlig sektor"?

Dette er et spørsmål til deg om å delta i et forskningsprosjekt hvor formålet er å se nærmere på KI-etikk og knytte det opp mot den hurtige implementeringen av KI-teknologi i offentlig sektor. I dette skrivet gir jeg deg informasjon om målene for prosjektet, og hva deltakelse vil innebære for deg.

Formål

Temaet for dette forskingsprosjektet er etisk implementering av kunstig intelligens i norsk offentlig sektor fra et humanistisk perspektiv. Oppgaven vil studere konsekvensene den teknologiske utviklingen har på enkeltmennesket og samfunnet.

Problemstillinger:

- Hvorfor følger Norge EU sine retningslinjer for etisk bruk av KI, er de gode nok?
- Trenger vi egne retningslinjer for å regulere teknologien? Hvis ja, hvorfor?
- Hvilke konsekvenser har det om vi ikke regulerer teknologien?

Hvem er ansvarlig for forskningsprosjektet?

Prosjektet tilhører Institutt for lingvistiske, litterære og estetiske studier ved Universitetet i Bergen. Ansvarlig for prosjektet er masterstudent Jenny Olsen Geithus. Veileder på prosjektet er dr. Ragnhild Solberg.

Hvorfor får du spørsmål om å delta?

Det er lite litteratur tilgjengelig som tar for seg KI-etikk opp mot norske problemstillinger og utfordringer. Du får derfor spørsmål om å delta i dette forskningsprosjektet på bakgrunn av din forskning innen fagfeltet KI.
Det er planlagt å gjennomføre 3-4 ulike intervjuer.

Hva innebærer det for deg å delta?

Hvis du velger å delta i dette forskingsprosjektet, innebærer det ett intervju som vil ha en varighet på ca. én time. Intervjuet vil følge en semistrukturert metode, men primært handle om ditt akademiske forhold til KI-etikk og dine tanker om norsk tilnærming og retningslinjer.

Personvern og samtykke

Det er frivillig å delta i prosjektet. Hvis du velger å delta, kan du når som helst trekke samtykket tilbake uten å oppgi noen grunn. Alle dine personopplysninger vil da bli slettet. Det vil ikke ha noen negative konsekvenser for deg hvis du ikke vil delta eller senere velger å trekke deg.

Jeg vil bare bruke opplysningene om deg til formålene jeg har fortalt om i dette skrivet. Jeg behandler opplysningene konfidensielt og i samsvar med personvernregelverket. Vi behandler opplysninger om deg basert på ditt samtykke. Prosjektet er godkjent av RETTE som er UiBs system for oversikt og kontroll med behandling av personopplysninger i forsknings- og studentprosjekter.

Alle opplysninger vil bli lagret på privat datamaskin med kodelås. Alle intervjuobjekt vil få muligheten til å anonymiseres om ønskelig. Om intervjuobjekt tillater det, vil navn, institusjon og tittel publiseres. Prosjektet vil etter planen avsluttes november 2024. Etter prosjektslutt vil datamaterialet med dine personopplysninger slettes.

Dine rettigheter

Så lenge du kan identifiseres i datamaterialet, har du rett til:

- innsyn i hvilke opplysninger vi behandler om deg, og å få utlevert en kopi av opplysningene
- å få rettet opplysninger om deg som er feil eller misvisende
- å få slettet personopplysninger om deg
- å sende klage til Datatilsynet om behandlingen av dine personopplysninger

Hvis du har spørsmål til studien, eller ønsker å vite mer om eller benytte deg av dine rettigheter, ta kontakt med:

- Institutt for lingvistiske, litterære og estetiske studier ved Universitetet i Bergen.
- Student: Jenny Olsen Geithus: *[fjernet for publisering]*
- Veileder: Ragnhild Solberg: *[fjernet for publisering]*

- Vårt personvernombud: [fjernet for publisering].

Hvis du har spørsmål knyttet til vurderingen som er gjort av personverntjenesten fra RETTE, ta kontakt med UIB's personvernombud: personvernombud@uib.no

Med vennlig hilsen
(Jenny Olsen Geithus)

Samtykkeerklæring

Jeg har mottatt og forstått informasjon om prosjektet *«etisk implementering av kunstig intelligens i norsk offentlig sektor»*, og har fått anledning til å stille spørsmål. Jeg samtykker til:

- ☐ å delta i semistrukturert intervju

- ☐ at opplysninger om meg – herunder navn, institusjon og tittel publiseres slik at jeg kan gjenkjennes
- ☐ at navn, institusjon og tittel ikke publiseres

- ☐ Jeg samtykker til at mine opplysninger behandles frem til prosjektet er avsluttet november 2024

(Signert av prosjektdeltaker, dato)