Handelshøyskolen BI - campus Bergen

BTH 36201

Bacheloroppgave - Økonomi og administrasjon

Bacheloroppgave

Ansvarsfordeling når kunstig intelligent militær-drone gjør fatal feil

Navn:	Tobias Berntsen

Utlevering:	06.01.2020 09.00
Innlevering:	03.06.2020 12.00

Ansvarsfordeling når kunstig intelligent militær-drone gjør fatal feil

Bacheloroppgave – Økonomi & Administrasjon

Eksamenskode: BTH 36201

6. semester

Handelshøyskolen BI Bergen

48 sider og 12000 ord (inkludert side i-vi og litteraturliste)

"Denne oppgaven er gjennomført som en del av studiet ved Handelshøyskolen BI. Dette innebærer ikke at Handelshøyskolen BI går god for de metoder som er anvendt, de resultater som er fremkommet, eller de konklusjoner som er trukket"

Sammendrag

Hovedformålet med denne studien var å besvare det overordnede forskningsspørsmålet:

«Hvor plasseres ansvaret når kunstig intelligent militær-drone gjør fatal feil?»

Syv hypoteser ble utviklet i tro på at tilstedeværelse av valgmuligheter, i vesentlig grad spiller en rolle for menneskers ansvarsplassering, slik det fremgikk i studiene til Cappelen et al., (2016) og Savani & Rattan (2012).

Gjennom et eksperiment, som hadde et mellomgruppe-design, ble to uavhengige grupper sammenlignet. Den ene gruppen ble forespeilet en situasjon hvor den menneskelige aktøren kunne velge å la det kunstig intelligente systemet ta beslutninger, eller, ta beslutninger selv. Den andre gruppen ble forespeilet en situasjon hvor dronen tok full-autonome beslutninger, og aktøren kunne derfor ikke påvirke dronens avgjørelser. Resultatene i studien viste at tilstedeværelse av valgmuligheter ikke medførte signifikante forskjeller i ansvarsfordelingen. Til tross for at deltakerne stilte seg mer kritisk til den menneskelige aktøren ved tilstedeværelse av valgmuligheter, var det ikke signifikante mellomgruppeforskjeller som impliserte at deltakerne i de to utvalgene hadde vesentlig forskjellige oppfatninger og holdninger til aktøren sin fremferd.

Funnene i studien viste at den norske stat i vesentlig grad pålegges et erstatningsansvar for den fatale feilen. I bredere sammenheng impliserer dette at stater/organisasjoner/selskaper bør være klar over at implementering av kunstig intelligent og autonom beslutningsteknologi, som kan volde sivil skade, kan medføre en betydelig ansvarliggjøring når disse gjør fatale feil – kanskje i enda større grad enn ved tradisjonell beslutningstaking, som i stor grad beror på den menneskelige aktøren sine evner.

Forord

Denne studien markerer slutten på et treårig bachelorstudium i økonomi og administrasjon, ved Handelshøyskolen BI Bergen.

Det har vært et lærerikt, spennende, men også krevende semester – på mange måter. Derfor vil jeg også rette en stor takk til BI som institusjon, og deres ansatte, for opprettholdelse av profesjonell undervisningspraksis denne våren.

Jeg ønsker å takke Mads Nordmo Arnestad for god veiledning, og for mulighet til fordypelse i et svært omfattende, intrikat og ikke minst viktig tema. Mitt engasjement for Big data, men også kunstig intelligens, samt tjenestegjøring i Forsvaret gjorde valget enkelt når denne muligheten åpenbarte seg. Studien har særlig vært spennende og motiverende fordi den omfatter en problemstilling, som per i dag ikke er reell, men som antagelig i nær fremtid vil være svært aktuell. Kunnskapen og erfaringen jeg sitter igjen med blir viktig for videre masterstudier.

Jeg konkluderte tidlig med at jeg ønsket at bacheloroppgaven skulle være et selvstendig verk, hvilket også har medført en vesentlig arbeidsmengde. Derfor vil jeg også takke familien for deres støtte, og deltakere i eksperimentet for verdifulle svar.

Tobias Berntsen
Bergen, våren 2020

Figurer og tabeller

Innholdsfortegnelse

Helt siden den første industrielle revolusjonen gjorde sitt inntog på slutten av 1700-tallet, har teknologiske nyvinninger gjort det mulig å erstattet mennesker med maskiner. Siden den gang har maskiner erstattet menneskers muskelkraft, men også utviklet seg til å utføre mer kompliserte og intelligente oppgaver. På 1950-tallet begynte forskere å se på muligheten for å skape kunstig intelligente (*Artificial intelligence, AI*) maskiner og systemer. Den verdensberømte britiske kryptologen og matematikeren, Alan Turing, hevdet at mennesker bruker tilgjengelig informasjon, så vel som fornuft, for å løse innfløkte problemer og ta beslutninger, så hvorfor kunne ikke maskiner gjøre det samme? I «*Computing Machinery and Intelligence*» skrev Turing om hvordan man kan bygge intelligente maskiner og teste deres intelligens, senere kjent som «Turing-testen» – en metode for å teste om en maskin er kapabel til å tenke som et menneske (Turing, 1950). I 1956 kom gjennombruddet da Allen Newell, Herbert A. Simon og Cliff Shaw utviklet det første kunstige intelligente dataprogrammet «Logic Theorists». De neste tiårene gikk utviklingen sakte, men fremover. De virkelige intelligente systemene så man først i slutten av det forrige århundre. Den kunstig intelligente maskinen «Deep Blue» forundret en hel verden i 1997 da den russiske stormesteren i sjakk Gary Kasparov måtte se seg slått av maskinen. Samtidig ble den første offentlig tilgjengelige programvaren for talegjenkjenning utviklet av Dragon Systems.

I dag er det flere som mener at vi beveger oss over i det som kan kalles «den fjerde industrielle revolusjonen», og kunstig intelligens er, om ikke den viktigste, i hvert fall en stor bidragsyter til «revolusjonen». Det kraftige fallet i produktivitetsvekst i Norge etter 2005 (fra 3 til 0,8 prosent årlig) var en viktig årsak til at regjeringen oppnevnte en produktivitetskommisjon (Jensen, 2016). Kommisjonen legger særlig vekt på teknologisk innovasjon som en betydelig driver for å snu den negative trenden, og å gå fra en ressursøkonomi til en kunnskapsøkonomi (NOU, 2016:3, s. 128). Regjeringen vil nå legge til rette for at Norge skal ha infrastruktur for kunstig intelligens i verdensklasse (Kommunal- og moderniseringsdepartementet, 2020, s.6). En McKinsey-studie estimerte at Norge har et automatiseringspotensiale med dagens teknologi på over 40 prosent (Chui, Manyika & Miremadi, 2017). PWC estimerte at globalt-BNP kan være 14% høyere i 2030 (tilsvarer 15,7 billioner USD) som følge av AI (Rao & Verweij, 2017).

I dagens moderniserte samfunn støter man daglig på kunstig intelligente og autonome systemer. Det være seg prissettingssystemer, selvstyrte kjøretøy, AI-styrt markedsførings- og annonsørinnhold eller chatbots. Det brutale koronaviruset «SARS-CoV-2», som fortsatt utgjør en reell trussel for verdenssamfunnet og -økonomien, ble oppdaget av en kunstig intelligens kalt «BlueDot», utviklet for å oppdage sykdomsutbrudd. Systemet baserer seg blant annet på språkanalyse og flyreisedata, og systemets algoritmer «scanner» blant annet offisielle rapporter og profesjonelle nettforum. Nøkkelord som «Lungebetennelse» og «Ukjent årsak» var utslagsgivende for konklusjonen om et virusutbrudd i Wuhan 31.desember 2019 (Aase, 2020). Dette var flere dager før *Centers for Disease Control and Prevention* og *World Health Organization* offentlig uttrykte bekymring for det nye viruset.

Kunstig intelligens skaper altså et mer produktivt og effektivt samfunn, og det er liten tvil om at denne vekstfaktoren vil få større betydning for verdenssamfunnet i fremtiden. I *Global Risks Report (2017)* betegnes kunstig intelligens som en av de fremvoksende teknologiene med størst nytteverdi, men også med størst skadepotensial (World Economic Forum, 2017). I denne sammenheng reiser det seg naturlig nok mange spørsmål og dilemmaer knyttet til ulike aspekter ved kunstig intelligens. Juridiske, etiske og moralske dilemmaer knyttet til AI er noen eksempler på aspekter som lenge har skapt grobunn for debatt.

Kanskje et av de mest omdiskutert etiske dilemmaene omhandler hvordan en AI skal opptre i tilfeller hvor ulykker inntreffer. I et tenkt scenario med et selvstyrt kjøretøy kan AI-en velge mellom å ta livet av en fotgjenger eller sjåføren selv. Hva er *riktig* eller *galt* i så henseende?

Et annet etisk dilemma handler om personvern og hvorvidt eksempelvis myndigheter eller selskaper skal kunne bruke AI til å gjøre inngripener i menneskers privatliv. Kina har i nyere tid blitt kritisert for å klassifisere enkeltindivider etter hvor lovlydige borgere de er. De har sosiale poengsystemer som belønner gode handlinger, mens dårlige handlinger straffes. Ved bruk av blant annet avanserte systemer for ansiktsgjenkjenning kan de registrere og kartlegge enkeltindivider, og store deler av befolkningen. Systemene fører til at individer presses til å være forsiktige med for eksempel hvor de går, hva de leser, og hvem de omgås. Over ti millioner mennesker i Kina har blitt nektet å reise med

fly eller tog på bakgrunn av hendelser som systemene har registrert (Almås, 2019).

«Big data» har i nyere tid vunnet frem som en stor inntekts- og påvirkningskilde for svært mange, med både etiske og uetiske formål. Virksomheter, myndigheter eller andre interessenter kan bruke elektroniske spor og informasjon, og særlig igjennom algoritmisk behandling av disse nå brukere målrettet, eksempelvis for å oppnå et politisk mål. «Cambridge Analytica»-skandalen er et eksempel på hvordan bearbeiding av persondata i stor skala kan ha enorme samfunnsmessige innvirkninger (Confessore, 2018). I kjølvannet av anvendelse av slike metoder, for å påvirke menneskers atferd og valg, har en ny term vokst frem – «Fake News». I forbindelse med spredning av falske nyheter har en AI kalt «Deepfake» blitt benyttet. «Deepfake» benytter seg av teknikker fra AI og maskinlæring for å manipulere eksisterende videoklipp. Personer kan byttes ut, det de uttrykker i videoen, enten verbalt eller kroppslig, kan manipuleres, og det ferdige produktet fremstår fortsatt ekte og troverdig.

I mars 2018 forekom verdens første offentlig kjente tilfelle av en dødelig ulykke mellom et selvstyrt kjøretøy og et menneske. Amerikanske Elaine Hertzberg skulle krysse en firefeltsvei i Arizona da hun ble påkjørt av et av Ubers selvstyrte testkjøretøy. Uber hadde siden august 2016 operert med selvstyrte kjøretøy i området, men testingen ble etter ulykken innstilt i nesten ett år. Etter hendelsen ble det rettet krass kritikk mot Uber og deres selvstyrte kjøretøy sine sikkerhetsmessige evner. *US National Transportation Safety Board* avdekket i etterkant en rekke feil i Uber sine systemer, deriblant mangler i programvaren, som ikke evnet å identifisere Hertzberg som en fotgjenger (Cuthbertson, 2019). Som ofte ellers i saker mellom enkeltpersoner og store selskaper endte det hele til slutt i et forlik mellom Uber og de etterlatte.

Det er følgelig mange viktige, vanskelig, og store spørsmål som reiser seg i forbindelse med utvikling av kunstig intelligens. Derfor er det heller ikke så rart at ulike beslutningsmyndigheter stadig utformer nye retningslinjer for å imøtekomme utviklingen. Kommunal- og moderniseringsdepartementet (2020) er tydelig på at kunstig intelligens i Norge skal bygge på etiske prinsipper, respekt for personvernet og god digital sikkerhet. Samtidig er det krevende å regulere en teknologi, som utvikler seg raskt. For tidlig regulering kan forme utviklingen på

en utilsiktet måte, skape skjevheter i markedet og begrense potensialet for innovasjon. (Kommunal- og moderniseringsdepartementet, 2020, s.21).

I forlengelsen av at kunstig intelligente og autonome systemer gjør feil vil denne studien undersøke menneskers oppfatninger og holdninger til ansvarsfordeling når AI gjør feil, med dødelig utfall. Hvor skal ansvaret plasseres når kunstig intelligente systemer, som er mer sofistikert enn menneskelig ekspertise, gjør fatale feil? Og har det faktum at man kan velge å avstå fra å benytte seg av AI-en noe å si for hvor mennesker plasserer ansvaret? Følgelig vil studien overordnede forskningsspørsmål være:

«Hvor plasseres ansvaret når kunstig intelligent militær-drone gjør fatal feil?»

På bakgrunn av oppgavens overordnede forskningsspørsmål formuleres syv hypoteser i den hensikt å besvare spørsmålet. Ved hjelp av en randomisert vignettundersøkelsen med et hypotetisk scenario, samt anvendelse av relevante teorier, er det et adekvat grunnlag for å kunne besvare problemstillingen. Likevel, det er svært få like og tilgjengelige studier med tilsvarende problemstillinger, rett og slett fordi det ikke anvendes slik sofistikert teknologi enda – innenfor fagområdet militære våpensystemer. Dette gjør oppgaven utfordrende, og det er motiverende å kunne skrive om noe, som antakelig, i nær fremtid, vil utgjøre en reell problemstilling.

1.0 Teoretisk rammeverk og hypoteseutvikling

1.1 Kunstig intelligens

Det finnes flerfoldige definisjoner av kunstig intelligens (KI), eller «*artificial intelligence*» (AI), og definisjonene endrer seg i takt med utviklingen, eller slik det beskrives i Teslers teorem: «Kunstig intelligens er det som enda ikke er gjort» (Hofstadter 1980, s. 601). Det brede spekteret av definisjoner gir ulike fortolkninger av hva kunstig intelligens faktisk er. I regjeringens *nasjonale strategi for kunstig intelligens* (2020), utformes definisjonen på bakgrunn av *European commission* (2019) sin ekspertgruppes tilnærming:

«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, s.9).

Systemer som baserer seg på kunstig intelligens kan enten tolke (eksempelvis ved mikrofoner, sensorer, kameraer etc.) eller innhente data fra andre kilder. Basert på dette vil systemene kunne analysere, ta beslutninger, og utføre handlinger. I enkelte AI-systemer finnes det i tillegg tilbakemeldingssløyfer, som gjør at AI-en lærer, kalt maskinlæring. Slik læring kan enten være erfaringsbasert, eller regelbasert ved tilbakemeldinger fra bruker. Det er som regel løsninger basert på maskinlæring som forbindes med kunstig intelligens (Kommunal- og moderniseringsdepartementet, 2020, s.11).

Stadig utvikling innen stordata («Big data») gjør det mulig å analysere store og komplekse datamengder mer effektivt og nøyaktig enn tidligere (Andersen & Bakkeli, 2015). Teknologien benyttes særlig av kunstig intelligente systemer til å skaffe et bredt analysegrunnlag. Dette gir mulighet for bedre prediksjoner og smartere beslutninger (McAfee & Brynjolfsson, 2012). Utvikling av stordata har også ført til økt anvendelse av algoritmer (Davenport & Harris, 2017). Maskinlæringsalgoritmer benyttes i hovedsak sammen med stordata, og baserer seg på matematiske modeller. Den kartlegger forhold mellom variabler som alminnelig menneskelig intelligens ikke evner å se. Dette skaper et bredere

innsyn, hvilket kan gi større utfordringer, men også muligheter for virksomheter (Coglianese & Lehr, 2016).

Kunstig intelligens, slik vi kjenner den i dag, kalles «smal»-AI og innebefatter systemer som er konstruert for å utføre én eller få spesifikke oppgaver. Ofte er slike AI-systemer komponenter i et større system. Det er denne typen AI som distribueres i dag, eksempelvis mønstergjenkjenning og bildebehandling. I dag finnes det ingen kunstig intelligens som ligner menneskelig intelligens (Kommunal- og moderniseringsdepartementet, 2020, s.9). Det er mange etiske, vitenskapelige, og teknologiske utfordringer knyttet til å utvikle en kunstig generell intelligens, som ligner menneskelig intelligens. Sunn fornuft, evnen til å resonere, selvinnsikt og maskinens evne til å definere egne formål er eksempler på utfordringer. (European Commission, 2019, s. 5).

1.2 Militære droner – kunstig intelligente og autonome systemer

Droner, ofte referert til som ubemannet luftfartøy («Unmanned Aerial Vehicle», UAV), krever ingen menneskelig interaksjon om bord i fartøyet. Hovedargumentene for anvendelse av slike fartøyer i militær sammenheng er at de kan utføre oppgaver som omtales som de 3 D-er: «Dull, dirty and dangerous» (Hexmoor, 2013, s.4). Militære droner har siden begynnelsen av 1990-tallet blitt brukt i konvensjonell krigføring. I første omgang ble droner brukt til overvåking og etterretning. Droner ble særlig brukt til slike formål under Kosovo-krigen (1998-1999) for å oppdage skjulte serbiske stillinger (Sabbagh, 2019). De første kjente tilfellene av våpenutstyrte droner så man etter 9/11 og USA sin «War on terror» i Irak og Afghanistan. Siden den gang har bruken av våpenutstyrte droner eksplodert. I perioden 2014-2018 sto droner for 42% av Storbritannia sine luftoppdrag mot IS, og 23% av luftangrepene (Drone Wars, 2012). *Royal United Service Institute* hevder at droner er fem til seks ganger så effektive som ved konvensjonelle luftoppdrag (Sabbagh, 2019). Utviklingen av militære droner har ført til at personell kan styre droner fra eksempelvis flyvåpenbaser i Lincolnshire, Storbritannia, og i Nevada, USA, uten å selv være eksponert i krigsområder i Midtøsten. Droner trenger fortsatt menneskelige støtte lokalt for avgang og landing.

I dag finnes det mange forskjellige typer militære droner, som brukes til mange ulike typer formål. De kanskje mest kjente militære dronene benyttet i offensive operasjoner er General Atomics sine «MQ-9 Reaper» og forløperen «MQ-1 Predator». Det amerikanske forsvaret og CIA har benyttet seg av disse siden tidlig 2000-tallet, og «MQ-9 Reaper» ble nylig benyttet av CIA i en planlagt likvidering av den iranske generalen Qassem Soleimani (Read, 2020). Det norske forsvaret er også ledende innen droneteknologi, og sammen med det norske selskapet Prox Dynamics utviklet de nanodronen «Black Hornet, PD-100 PRS», som veier 17,5 gram og anvendes i overvåkings- og etterretningsoperasjoner (Arstad, 2019).

Droner som tar autonome avfyringsbeslutninger, basert på en sofistikert kunstig intelligens, uten at menneskelige vurdering ligger til grunn, har ikke blitt brukt i krigføring enda. Dette hevdes blant annet i Paul Scharre's (forsvarsekspert i Pentagon) prisvinnende bok «*Army of none: Autonomous Weapons and the Future of War*», og i Dyndal, Berntsen & Redse-Johansen (2017) sin rapport i *NATO review*, samt flere omfattende droneartikler, deriblant Doyle (2018) og Sabbagh (2019). Mange frykter slike våpensystemer og bruken av dem i fremtidig krigføring, samtidig som det for ethvert forsvar vil være uvurderlig ressurs å besitte. Stephen Hawking hevdet at AI kan være det største menneskelige gjennombruddet i historien, men også det siste, dersom vi ikke mestrer å bruke den riktig (Hawking, 2014). I 2015 gikk hundrevis av forskere og teknologer, med blant annet Stephen Hawking og Elon Musk i spissen, ut og advarte mot potensielle trusler ved utvikling av AI (Griffin, 2015). I et «open letter» signert av representanter fra Google, Deep Mind, og noen av de største universitetene i USA (Cambridge, Oxford, Harvard, Stanford og MIT), poengteres viktigheten av forskning på AI som kan utgjøre en trussel for menneskeheten (Russell, Dewey & Tegmark, 2015).

Utvikling av AI-styrte droner og autonome våpensystemer har lenge blitt kritisert for å være uetisk og potensielt farlig. Nevnte Hawking og Musk, samt Steve Wozniak, har vært ute og frarådet bruk og utvikling av autonome og kunstig intelligente våpensystemer (Gibbs, 2015). Det er ikke bare formålet og i hvilke hender slike systemer havner som vekker bekymring hos forskere, men også tanken på at et «AI-kappløp» fører til at militærmakter for tidlig utplasserer AI-systemer, som er underutviklet og kan volde stor skade (Scharre, 2020). I tillegg bekymres flere over at «liv-eller-død»-avgjørelser skal flagges ut til autonome

våpensystemer, at dette undergraver verdien av menneskeliv, og således er umoralsk. FNs generalsekretær António Guterres mener at utvikling av maskiner som kan ta menneskeliv, uten menneskelig intervensjon, er politisk uakseptabelt og moralsk forkastelig, og at dette bør forbys gjennom Folkeretten (UN, 2019). Dyndal et al., (2017) hevder at det kan argumenteres for at bruk av autonome droner kan aksepteres, ikke bare i et moralsk perspektiv, men at det i tillegg kan være moralsk gunstig. Begrunnelsen for dette er blant annet at droner kan prosessere vesentlig mer informasjon enn mennesker, og derfor ta velbegrunnede beslutninger. Droner påvirkes heller ikke av støy eller følelser, hvilket kan redusere risikoen for krigsforbrytelser.

1.3 «Algoritme-aversjon»

Siden tidlig 1950-tallet har statistiske algoritmer beviselig vært mer treffsikre enn mennesker på flere områder (Meehl, 1954; Sawyer, 1966; Einhorn, 1972; Dawes, 1979; Dawes, Faust & Meehl, 1989). I nyere tid er det flere eksempler på at mer sofistikert teknologi, som kunstig intelligens og maskinlæring, er mer treffsikre enn mennesker. En studie viste at en AI-algoritme var mer nøyaktige i predikering av hvilke innsatte som kom til å fortsette med kriminelle lovbrudd (Temming, 2020). En annen AI-algoritme gjorde riktige vurderinger i over 90 prosent av tilfellene, mot legenes 77,5 prosent i triageprosesser (Donnelly, 2017). IBM sin Watson-maskin, som baserer seg på maskinlæring, ble sammenlignet med medisinsk ekspertise på tvers av 1000 kreftdiagnoser. I 30 prosent av tilfellene fant Watson-maskinen behandlingsformer som ekspertene gikk glipp av (Lohr, 2016).

Til tross for at relativt simple statistiske algoritmer er mer treffsikre enn menneskelig ekspertise på mange områder, er det i dag en systematisk tendens til at mennesker ikke ønsker å benytte seg av disse i vurderings- og beslutningsprosesser. Denne psykologiske motstanden, kalt «algoritme-aversjon» (Dietvorst, Simmons & Massey, 2014), viser seg i mange aspekter i samfunnet. Enten det gjelder å selektere arbeidssøkere, stille kliniske diagnoser i helsevesenet eller ta økonomiske og strategiske beslutninger, så favoriseres menneskelige vurderinger og beslutninger.

Dietvorst et al., (2014) ønsket i en studie å undersøke fenomenet «algoritme-aversjon», og fant at mennesker systematisk unngår bruk av algoritmer i prediksjoner. I en del av studien ble deltakere presentert for inntaksdata for et masterstudium, og på bakgrunn av dette skulle det predikeres hvilken karakter de ulike studentene ble uteksaminert med. Deltakerne kunne velge mellom å satse penger på egen prediksjon, en annen menneskelig prediksjon, eller en statistisk algoritmisk prediksjon. Når deltakerne fikk observere algoritmens tidligere prediksjoner, og så at den gjorde feil ved noen tilfeller, ble de mindre tilbøyelige til å velge algoritmen. Kanskje ikke så «algoritme-averst» med tanke på at den faktisk bommet ved noen tilfeller? Forbausende nok valgte deltakerne fortsatt menneskelige prognosemakere selv når de fikk observere at algoritmen var mer treffsikker. Resultatene impliserer at vi stoler mindre på maskinelle enn menneskelige vurderinger, og bekrefter således at menneskelige vurderinger og beslutninger favoriseres.

I en oppfølgingsstudie ønsket Dietvorst et al., (2016) å undersøke om det var mulig å redusere «algoritme-aversjonen» ved å tillate mindre justeringer på den algoritmiske prediksjonen. Resultatet viste at selv minimal mulighet for påvirkning på prediksjonen innebar en større villighet til å velge algoritmen blant deltakerne. Resultatene viste altså at tilliten til algoritmiske prognoser øker ved mulighet for noe brukermedvirkning.

1.4 Moralsk ansvar

En viktig forskjell mellom mennesker og andre skapninger er at kun mennesker kan være moralsk ansvarlig for hva de gjør (Fischer & Ravizza, 1998, s.1). Fischer & Ravizza (1998) illustrer dette ved et hypotetisk scenario hvor en finner vasen sin ødelagt. De hevder at man vil reagere forskjellig om ødeleggelsen skyldes en gjest som med forsett knuste vasen, i motsetning til om eierens hund hadde vært uheldig. En blir kanskje skuffet, irritert og sint på hunden, men man vil ikke føle en like sterk moralsk indignasjon som overfor gjesten som knuste vasen med vilje. Til tross for at både gjesten og hunden er kausalt ansvarlig for hendelsen vil man holde gjesten ansvarlig i en helt annen forstand enn hunden. Begge er *ansvarlig* for utfallet, men kun gjesten kan være *moralsk ansvarlig* (Fischer & Ravizza, 1998, s.2).

For at et individ skal kunne være moralsk ansvarlig for sine handlinger legger Fischer & Ravizza (1998) to forhold til grunn. Det ene forholdet, som de kaller «epistemic condition», handler om at et individ kjenner til de faktiske forholdene som omgir handlingen, og handler med den rette typen tro og intensjoner. Det andre forholdet, som de kaller «control condition», handler om at individet må ha kontroll over sin atferd «i en passende forstand». Gitt denne tilnærmingen, må individet kjenne til de faktiske omstendighetene for sitt valg, fritt kunne ta en beslutning, og velge en passende handling basert på disse omstendighetene, for å skulle holdes moralsk ansvarlig.

Gray, Young & Waytz (2012) la særlig vekt på intensjonen bak handlinger i tilknytning moralsk ansvar i artikkelen «*Mind Perception is The Essence of Morality*». De hevdet at intensjonen bak handlinger er så kraftig tilknyttet skylden, at selv irrelevante intensjoner kan øke skyld- og ansvarsspørsmålet. Denne tilnærmingen bekreftes blant annet i et eksperiment utført av Woolfolk, Doris & Darley (2006), hvor mennesker som tvinges til å skyte andre, oppfattes som mer umoralske når de faktisk ønsker at vedkommende skal dø.

Filosofen Peter French hevdet at moralsk ansvar ikke bare kan tildeles individer, men også organisasjoner. Ettersom organisasjoner absorberer og utstråler menneskers intensjoner og handlinger, er beslutningene derfor også organisatoriske. French mente at organisasjoners intensjoner er nok til at organisasjonens handlinger blir gjort med «vilje», og dermed skal organisasjonen også holdes moralsk ansvarlig (Shaw & Barry, 2015).

Institutt for kunstig intelligens (CAIR) ved Universitet i Agder forsker på å gi kunstig intelligens et moralsk kompass. Professor ved CAIR, Einar Duenger Bøhn, hevdet at kunstig intelligens per i dag ikke har en egen moral, men at den kan simuleres (Tolfsen, 2018). Eksempelvis vil iPhone sin «Siri»-funksjon respondere høflig fordi den er programmert til det. Bøhn mente at kunstig intelligente systemer ikke kan ta moralske valg selv, og at utviklingen av kunstig moral ikke har kommet lengere siden 1950-tallet (Tolfsen, 2018).

Tradisjonelt holdes produsenter og/eller operatører moralsk og legalt ansvarlig for feil begått av maskiner (Matthias, 2003). Anvendelse av algoritmisk beslutningsteknologi innebærer likevel en dimensjon som har formet en ny problemstilling. Når maskiner baserer seg på algoritmiske beslutninger, som

produsenten, organisasjonen eller operatøren av maskinen i prinsippet ikke er i stand til å forutse, blir det ikke lenger entydig hvor ansvaret skal plasseres. Hvem skal holdes ansvarlig når kunstig intelligens ikke kan være moralsk ansvarlig for sine handlinger? Denne situasjonen refereres ofte til som «The black-box problem» (Bathee, 2018), og skaper det flere omtaler som et «responsibility gap» (Matthias, 2003; Dyndal et al., 2017).

1.5 Personlig ansvar

Prinsippet om at mennesker bør holdes personlig ansvarlig som følge av de konsekvenser ens valg medfører, er et fundamentalt moralsk ideal i det vestlige samfunn (Cappelen, Fest, Sørensen & Tungodden, 2016). Dog har tolkning og anvendelse av dette overordnede prinsippet i lang tid blitt debattert. Det finnes flere eksempler som implisitt gir uttrykk for at politikk i visse tilfeller baserer seg på personlig ansvar. Amerikanske myndigheters reduksjon i overføringer til enslige forsørgere og familier med arbeidsledige ser ut til være forankret i antakelsen om at disse gruppene bør holdes personlig ansvarlig for sin egen situasjon (Moffitt, 2015). Likeledes argumenteres det for at den politiske diskursen knyttet til livsstilsrelaterte sykdommer forstås ved prinsippet om personlig ansvar (Wikler, 2002; Brownell et al., 2010).

Det er naturligvis flere nyanser ved et valg, som gjør prinsippet om personlig ansvar noe generaliserende, og ikke minst situasjonsbetinget. Særlig når flere parter involveres, eksempelvis en AI, kan situasjonen bli langt mer komplisert. Hvor stor del av utfallet kan relateres til aktøren som samhandler med AI-en? Hvor stor brukermedvirkning har aktøren? Er valget delvis frem-/påtvunget?

Ifølge filosofen Peter Vallentyne (2008) skal ikke individer holdes personlig ansvarlig for utfall av ens valg, dersom:

1. *Personen kunne ikke ha endret sannsynligheten for utfallet ved å velge annerledes*, eller
2. *Personen kunne bare avverget utfallet ved en urimelig stor kostnad/ulempe*. (Ikke akseptabelt alternativ)

Vallentyne (2008) omtaler disse betingelsene som «minimale betingelser» som må være tilstede for at individer ikke skal holdes personlig ansvarlig. Valget et

individ foretar seg er således «moralsk relevant» dersom det ikke bryter med betingelsene. Begge betingelsene fanger opp situasjoner hvor individer på forhånd ikke har noen grunn til å velge annerledes. Det kan derfor diskuteres hvorvidt individer faktisk har utøvd selvstendige og frie valg, kalt autonomi i moralfilosofien.

Cappelen et al., (2016) ønsket i sin studie å undersøke om mennesker holder individer personlig ansvarlig og likevel bryter med nevnte betingelser. De fant signifikante bevis for at disse betingelsene brytes. I studien skulle en tredjepart fordele inntekter mellom to deltakere. I første del av studien hadde deltakerne ingen valg. Inntjeningen deres ble utelukkende bestemt av et tilfeldig lotteri, hvor en av deltakerne «vant». På bakgrunn av dette kunne tredjeparten fritt velge å omfordele gevinsten mellom deltakerne. I den andre delen av studien hadde deltakerne valgmuligheter på forhånd:

1. I det ene tilfellet hadde deltakerne valget mellom to lotterier, som var identiske på forhånd, de var dermed ikke i stand til å endre sannsynligheten for utfallet (betingelse 1).
2. I det andre tilfellet hadde deltakerne valget mellom et lotteri og et «trygt alternativ». Det trygge alternativet fremsto ikke som et akseptabelt alternativ, men en betydelig ulempe sammenlignet med gevinsten man kunne oppnå i lotteriet (25 NOK utbetalt mot lotteriet som hadde en forventningsverdi på 400 NOK) (betingelse 2).

Dersom tredjepartene hadde støttet de minimale betingelsene, burde ikke innføring av valgmuligheter, i de to siste scenarioene, påvirket omfordelingen av gevinster. Det burde således være samme nivå av inntektsulikhet omfordelt av tredjeparten i første del av studien, som i andre del av studien. Studien fant at dette ikke var tilfelle. Tilstedeværelse av valgmuligheter førte til en stor økning i villighet til å akseptere inntektsulikhet mellom deltakerne. Tredjepartene overførte mindre til deltakerne uten gevinst når disse hadde valgmuligheter. Dette betyr at tredjeparten mener at deltakerne som kom verst ut, var personlig ansvarlig for utfallet, og fortjente det, i situasjoner hvor de egentlig ikke hadde grunn til å velge annerledes. I likhet med Cappelen et al., (2016) sin studie undersøkte Savani & Rattan (2012) dette «valgfenomenet». Deres funn viste at fremheving av valgbegrepet, som er høyt verdsatt i USA, gjorde at mennesker rettferdiggjør

økonomisk ulikhet, fordi mennesker legger større vekt på individuelle valg enn sosiale og samfunnsmessige faktorer, for utfall av menneskers liv.

1.6 Hypoteser

Gjennom presentert teoretiske rammeverk fremgår det at helt enkle statistiske algoritmer, og mer sofistikert teknologi som AI og maskinlæring, i mange tilfeller er mer treffsikre og effektive enn menneskelig eksperter (Meehl, 1954; Sawyer, 1966; Einhorn, 1972; Dawes, 1979; Dawes, Faust & Meehl, 1989; Lohr, 2016; Donnelly, 2017; Temming, 2020). Dietvorst et al., (2014) fant likevel at mennesker favoriserer menneskelige beslutningstakere foran mer treffsikre statistiske algoritmer.

Det er ingen troverdige empiriske studier som konkluderer med at det i dag anvendes kunstig intelligente og full-autonome militære droner. Det er likevel konsensus om at slike droner er under utvikling, og at de i fremtiden kan bli viktige i moderne krigføring (Dyndal et al., 2017; Doyle, 2018; Scharre, 2018; Sabbagh, 2019; Scharre, 2020). Når det likevel ligger til grunn i denne undersøkelsen at dronen benytter seg av en kunstig intelligent algoritme i beslutningstaking, som er mer sofistikert enn menneskelige droneførere, skapes det flere omtaler som et «responsibility gap». Tradisjonelt holdes produsenter og/eller operatører moralsk og legalt ansvarlig for feil begått av maskiner. Situasjonen blir mindre entydig når maskiner baserer seg på algoritmiske beslutninger, som produsenten/operatøren av maskinen i prinsippet ikke er i stand til å forutse (Matthias, 2003; Dyndal et al., 2017).

Fischer & Ravizza (1998) legger til grunn i sin forklaring på moralsk ansvar at kun mennesker kan være moralsk ansvarlig for sine handlinger. Samtidig understreker professor Einar Duenger Bøhn at kunstig intelligens, per i dag, ikke kan ha en egen moral (Tolfsen, 2018). Filosofen Peter French hevder at organisasjoner også kan holdes moralsk ansvarlig (Shaw & Barry, 2015).

Cappelen et al., (2016) og Savani & Rattan (2012) viste i sine studier at mennesker er mer tilbøyelige til å plassere ansvar dersom det eksisterer valgmuligheter. Hypotesene i denne studien bygger på denne konklusjonen. Hvor mye påvirker tilstedeværelse av valgmuligheter menneskers opplevde

kritikkverdighet, moralsk ansvarsfordeling, sinne, og ønsket om et straffe- og erstatningsansvar når AI gjør fatal feil?

På bakgrunn av det teoretiske rammeverket og eksperimentets design, er det utviklet syv hypoteser for å besvare studiens overordnede forskningsspørsmål:

H1: Respondentene mener at Sigurd i større grad fortjener kritikk når han har et valg (mulighet for manuell overstyring).

H2: Respondentene mener at Sigurd i større grad er moralsk ansvarlig når han har et valg (mulighet for manuell overstyring).

H3: Respondentene mener at Sigurd i større grad fortjener straff når han har et valg (mulighet for manuell overstyring).

H4: Respondentene føler i større grad på et sinne ovenfor Sigurd når han har et valg (mulighet for manuell overstyring).

H5: Respondentene mener at Sigurd i større grad bør pålegges et erstatningsansvar når han har et valg (mulighet for manuell overstyring).

H6: Respondentene mener at den norske stat i mindre grad bør pålegges et erstatningsansvar når Sigurd har et valg (mulighet for manuell overstyring).

H7: Respondentene mener at selskapet som lager den kunstige intelligente dronen i mindre grad bør pålegges et erstatningsansvar når Sigurd har et valg (mulighet for manuell overstyring).

2.0 Metode og forskningsdesign

I denne delen av oppgaven belyses benyttet fremgangsmåte for rekruttering og innsamling av data, samt eksperimentets design. Metode dreier seg om hvordan vi innhenter, organiserer og tolker informasjon (Larsen 2017, s.17).

For å besvare det overordnede forskningsspørsmålet benyttes en kvantitativ metode-tilnærming. Kvantitative metoder befatter seg med tall og det som er målbart, og skiller seg fra kvalitativ metode, som ser på egenskaper og meninger rundt fenomener. Undersøkelsen har et kausal design i form av et eksperiment, hvor hovedmålet var å avdekke hvilken effekt manipulasjon av den uavhengige variabelen hadde på de avhengige variablene. Det er naturlig at det overordnede forskningsspørsmålet og hypotesene til en viss grad styrer metodevalget. Siden studien var av eksperimentell art og bygger på individers respons på en vignett, var det formålstjenlig å prøve å kvantifisere dataene for å avdekke mønstre, som kunne besvare hypotesene og forskningsspørsmålet. En del respondenter uttrykte et ønske om å besvare en del av påstandene i undersøkelsen utover forhåndsdefinert graderingsskala. Ettersom formålet med studien i vesentlig grad handlet om å forsøke å avdekke mellomgruppeforskjeller for å besvare hypotesene, og forskningsspørsmålet, var det heller ikke hensiktsmessig å benytte en metodetriangulering, og elaborere rundt enkeltindividers meninger.

Studien består av en hybrid mellom primær- og sekundærdata, med hovedvekt på førstnevnte. Primærdata er data innsamlet til eget formål, gjennom vignettundersøkelsen, for å besvare hypotesene og problemstillingen. Sekundærdata er eksisterende data. Vitenskapelige forskningsartikler og rapporter tilknyttet tematikken i studien, som kunstig intelligens, algoritmisk beslutningsteknologi, moralsk ansvar og ansvarsfordeling, ble innhentet for å gi studien empirisk tyngde.

2.1 Validitet og reliabilitet

Validitet, eller gyldighet, dreier seg om hvor godt man måler det man har til hensikt å måle. Selv om et mål har høy reliabilitet, er det ikke sikkert at validiteten er høy (Gripsrud, Olsson & Silkoset, 2016, s. 61). I dette tilfellet beror validiteten på i hvilken grad en kan trekke gyldige slutninger på bakgrunn av

undersøkelsen. Reliabilitet handler om i hvilken grad vi kan stole på dataene, og hvor godt de representerer det aktuelle fenomenet som skal undersøkes. Dette betyr at dersom vignettundersøkelsen replikeres, skal resultatet bli det samme. For å styrke reliabiliteten til dataene formuleres enkelte påstander flere ganger, men med omformuleringer. Samtlige respondenter ble innledningsvis opplyst om at svarene deres behandles med absolutt anonymitet, slik at svarene i størst mulig grad baserer seg på ærlighet.

2.2 Rekruttering og utvalg

Undersøkelsen ble gjennomført i den elektroniske programvaren «Qualtrics», som er av de mest anerkjente programvarene for kvantitative statistiske undersøkelser. Rekrutteringen foregikk elektronisk. Invitasjoner ble primært sendt til et segment bestående av venner og bekjente. Som nevnt var undersøkelsen anonym, og deltakerne ble informert om at de fritt kunne velge å samtykke til deltakelse i tråd med regler for personvern i forskningsprosjekter (De Nasjonale Forskningsetiske komiteene, 2006). Om lag 500 personer ble invitert til deltakelse i undersøkelsen, primært via E-post, og de sosiale nettverkene Facebook og LinkedIn. Antall respondenter totalt var 340, med et frafall på 128 på det meste, hvorav demografiske spørsmål i undersøkelsens sluttfase var utslagsgivende for det meste av frafallet. Undersøkelsen ble ikke publisert i sosiale medier grunnet erfaringer med stort frafall, samt en oppfatning av at svarene bærer preg av å være mer pålitelige ved personlig invitasjon/forespørsel. Denne tilnærmingen var også hensiktsmessig ettersom det ble enklere å sikre en jevn fordeling i undersøkelsens demografiske data, og unngå stor utvalgs- og frafallsskjevhet. Respondentene ble kontaktet personlig og anmodet om å besvare undersøkelsen via tilsendt URL-lenke. Der ble de presentert for studien:

«Takk for at du ville delta i denne spørreundersøkelsen, utført av studenter ved Handelshøyskolen BI. Ansvarlig for studien er Mads Nordmo Arnestad. Formålet med studien er å undersøke hva folk tenker om kunstig intelligens og ansvar. Studien tar ca. 5 minutter å gjennomføre. Alle data som samles inn anonymiseres. Vi vil ikke kunne identifisere enkeltdeltakere. Du kan når som helst trekke deg fra studien. Deltakelse i studien medfører ingen fare for psykisk eller fysisk helse. Vennligst indiker ditt samtykke til å delta, og klikk videre.»

2.3 Eksperimentets design

Studiens undersøkelse er av typen «vignette survey». I vignett-metoden presenteres respondenter for korte historier av hypotetisk karakter, som de skal gjøre seg opp en mening om. I dette tilfellet to «identiske» historier, men med en manipulert variabel. Eksperimentet betegnes som et mellomgruppe-design, hvor to uavhengige grupper sammenlignes.

I første del av undersøkelsen ble deltakerne presentert for den hypotetiske historien om Sigurd Svendsen, som var spesialist i Luftforsvarets droneavdeling i 2025, hans virke som dronefører i kampen mot terrorisme, og den militære dronen, som tok avfyringsbeslutninger på bakgrunn av en sofistikert kunstig intelligens («Seektodecide3000»). Gjennom ansiktsgjenkjenning og verifisering i et register med kjente terrorister avgjør dronen om målet er legitimt. Hovedbegrunnelsen for å benytte en kunstig intelligens til å ta avfyringsbeslutninger var ønsket om å minimere feil. Undersøkelser viste at «Seektodecide3000» hadde lavere feilmargin, var mer treffsikker og effektiv enn menneskelig droneførere. Videre ble deltakerne presentert for kjernen i historien, nemlig at dronen feilaktig avfyrer en rakett og dreper en uskyldig lokal bonde, som den trodde var en terrorist, og bonden etterlot seg en enke. Til slutt i vignetten ble deltakerne presentert for undersøkelsens manipulerte variabel:

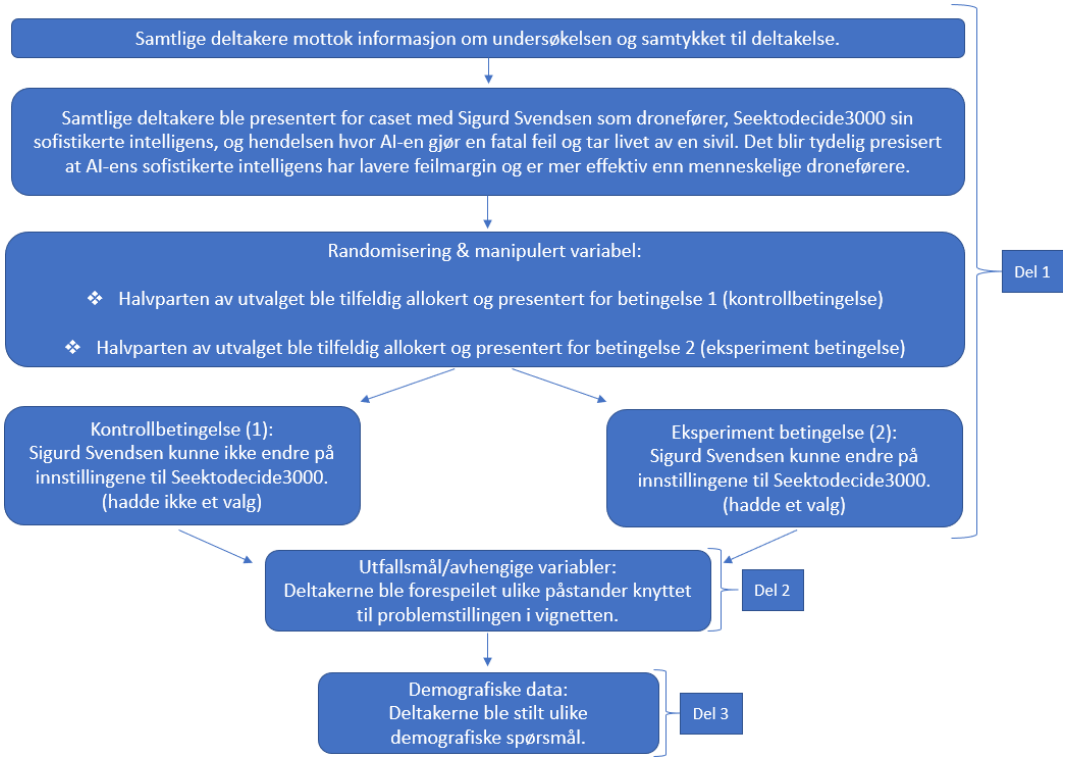
Sigurd Svendsen kunne ikke endre på innstillingene til «Seektodecide3000», den tok autonome avfyringsbeslutninger **(Sigurd hadde ikke et valg)**, eller, Sigurd Svendsen sto fritt til å endre innstillingene til «Seektodecide3000» og dens avfyringsbeslutninger **(Sigurd hadde et valg).**

I tilfellet hvor Sigurd hadde mulighet til å endre innstillingene, og selv ta kontroll, valgte han å la være. Halvparten ble presentert for kontrollbetingelsen (Sigurd hadde ikke et valg), mens den andre halvparten ble presentert for eksperiment betingelsen (Sigurd hadde et valg). Det er viktig å poengtere at deltakerne ble tilfeldig allokert og presentert for én av de to betingelsene.

I den andre delen av undersøkelsen skulle deltakerne besvare 14 påstander knyttet til problemstillingen i vignetten. For å besvare påstandene skulle respondentene indikere hvor uenig eller enig de var i påstandene, på en skala fra 1-7, der 1= helt uenig og 7= helt enig. Videre skulle deltakerne besvare to kontrollspørsmål knyttet til vignetten. Det første kontrollspørsmålet ble stilt for å sikre at

respondentene var kjent med betingelsen de ble forespeilet; enten hadde Sigurd et valg, eller, så hadde han ikke et valg. Det andre kontrollspørsmålet handlet om å sikre at respondentene var kjent med at menneskelige droneførere bruker lengere tid og gjør flere feil enn «Seektodecide3000». Dersom respondentene svarte feil på et eller begge kontrollspørsmålene, eller indikerte at de ikke visste/husket, ble de ekskludert for videre analyse i hypotesetesting.

I undersøkelsens siste del skulle deltakerne besvare seks demografiske spørsmål knyttet til: alder, kjønn, årslønn, arbeidssektor, lederroller, og høyeste fullførte utdannelse. I figuren under illustreres eksperimentets prosedyre, inndelt i de tre fasene.



Figur 1. Oversikt over eksperimentets prosedyre

2.4 Utfallsmål

I eksperimentet ble 7 ulike utfallsmål/avhengige variabler anvendt: «kritikkverdighet», «moralsk ansvar», «straff», «sinne», «erstatning fra Sigurd», «erstatning fra den norske stat», og «erstatning fra AI-selskapet».

De tre første avhengige variablene består av tre påstander, mens «sinne» består av to påstander. Påstandene hadde ulik formulering, men samme meningsinnhold, og utgjorde i sum de avhengige variablene. Dette ble gjort for å kvalitetssikre at deltakerne forsto de ulike påstandene, skape nyanserte svar, og som nevnt sikre reliable målinger. Ettersom erstatningsspørsmålene var av ulik karakter utgjorde de alene hver sin avhengige variabel.

En Likert-skala ble benyttet for å måle respondentenes holdninger, meninger og oppfatninger knyttet til påstandene. Intervallet i graderingsskalaen var fra 1 til 7, hvor 1= helt uenig, 2= uenig, 3= litt uenig, 4= nøytral, 5= litt enig, 6= enig, og 7= helt enig.

Den manipulerte variabelen; «Sigurd hadde et valg»/«Sigurd hadde ikke et valg». var eksperimentets uavhengige variabel.

3.0 Resultater

Alle resultater fra eksperimentet er behandlet i IBM SPSS. Først vil undersøkelsens demografiske sammensetning beskrives, dernest avdekkes sammenheng mellom dataene, og til slutt benyttes en uavhengig t-test til hypotesetesting for å avdekke mellomgruppeforskjeller.

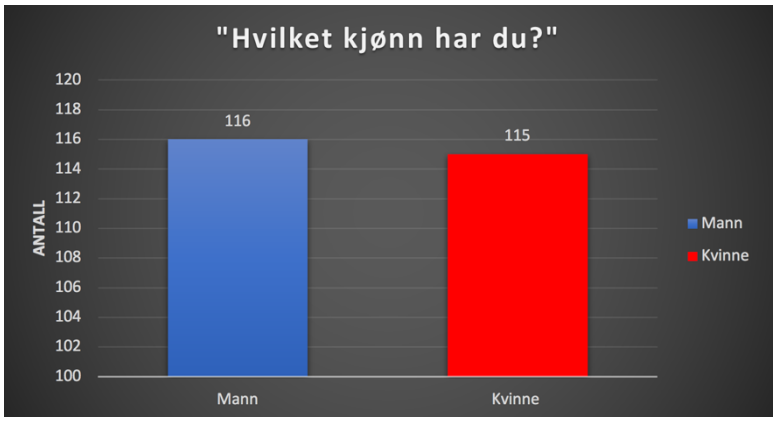
3.1 Deskriptive data

Det var totalt 340 som godtok å gjennomføre undersøkelsen, og 212 besvarte alle spørsmål. I undersøkelsens avsluttende del var frafallet størst. Spørsmålet om respondentene var leder på sin arbeidsplass eller ikke hadde størst frafall. 128 av 340 valgte ikke å besvarte spørsmålet.

Gjennomsnittsalderen til respondentene var 50,13 år, hvorav den yngste var 17 år, og den eldste var 83 år. Standardavviket var på 12,94, som indikerer at undersøkelsen har stor variasjon med hensyn til respondentenes alder.

Kjønn

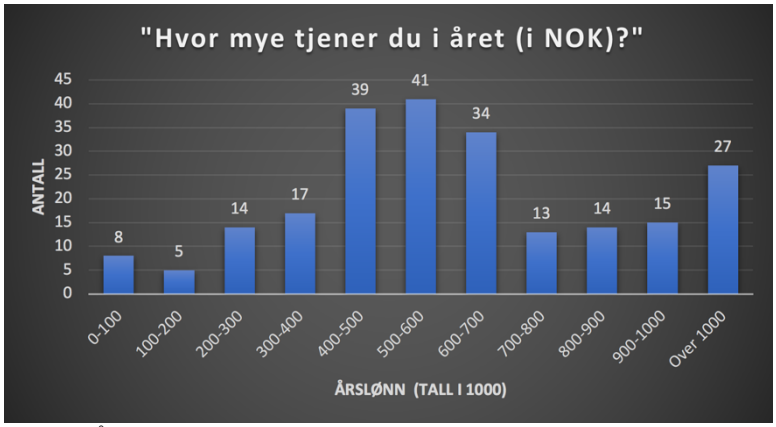
231 respondenter besvarte dette spørsmålet. Av disse var 116 (50,2%) menn, og 115 (49,8%) kvinner, hvilket harmoner med populasjonen for øvrig.



Figur 2. Kjønn

Lønn

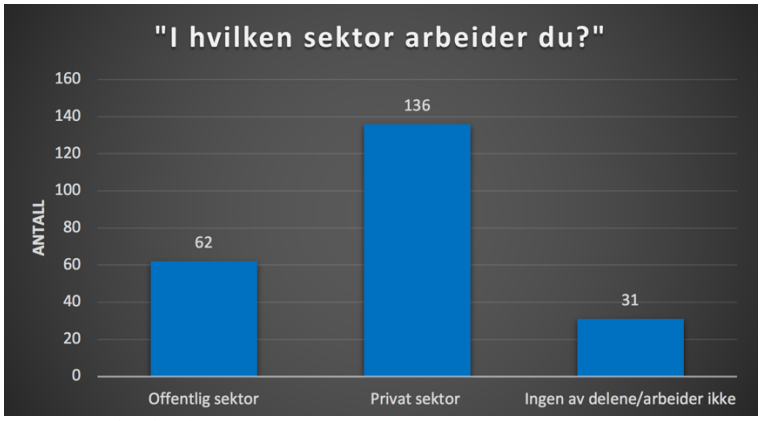
227 respondenter besvarte dette spørsmålet. Det er stor variasjon i årslønnen til respondentene, fra 0 NOK til over 1.000.000 NOK i året. Inntektsfordelingen er sentrert ved 400.000-700.000 NOK (>50%), med et flertall på 500.000-600.000 NOK. Gjennomsnittsårslønnen i Norge i 2019 var 570.800 NOK (Statistisk sentralbyrå, 2019). Følgelig gjenspeiler inntektsfordelingen populasjonen i Norge ganske godt.



Figur 3. Årslønn

Arbeidssektor

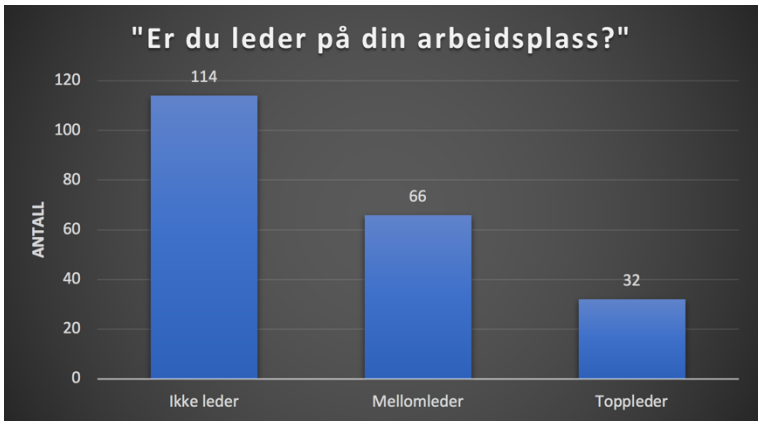
229 respondenter besvarte dette spørsmålet. 62 respondenter (27,1%) arbeidet i offentlig sektor, 136 respondenter (59,4%) i privat sektor, mens 31 respondenter (13,5%) ikke arbeidet. Spredningen mellom respondenter som arbeidet i offentlig og privat sektor harmonerer ganske godt med den norske arbeidsstyrken, hvor om lag 66% arbeider i privat sektor og 34% i offentlig sektor (SSB, 2020). Forholdet mellom respondenter som arbeider i offentlig og privat sektor gjengir således et nokså riktig bilde av fordelingen i populasjonen.



Figur 4. Arbeidssektor

Lederstilling

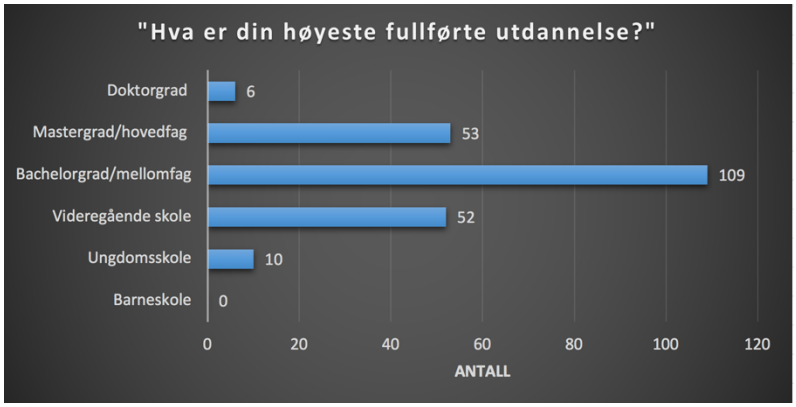
212 besvarte dette spørsmålet. 114 respondenter (53,8%) hadde ikke lederroller på sin arbeidsplass. 66 respondenter (31,1%) var mellomledere, mens 32 respondenter (15,1%) var toppledere.



Figur 5. Lederroller

Utdanning

230 besvarte dette spørsmålet. 10 respondenter (4,3%) svarte at deres høyeste fullførte utdannelse var ungdomsskolen, 52 respondenter (22,6%) svarte videregående skole, 109 respondenter (47,4%) svarte bachelorutdanning, 53 respondenter (23,0%) svarte masterutdanning, mens 6 respondenter (2,6%) svarte doktorgradsutdanning. Svært mange respondenter hadde høyere utdanning. Særlig bachelor-, master- og doktorgradsutdanning avviker en del fra populasjonen, som ligger på henholdsvis 22%, 7,3% og 0,7% (SSB, 2019).



Figur 6. Utdanning

3.2 Sammenheng mellom data

Cronbach's alfa

For å måle reliabiliteten mellom indikatorer som tilhører samme variabel benyttes Cronbach's alfa. En tommelfingerregel som brukes, er at verdien på Cronbach's alfa skal være større enn 0,7, men ikke for nær 1 dersom et multippelt mål skal beregnes (Gripsrud et.al, 2010). I beregning av Cronbach's alfa måtte Likert-skalaen til tre påstander reverseres for å få riktige alfa-verdier. Dette gjaldt påstandene knyttet til om Sigurd var moralsk ansvarlig, og om Sigurd fortjente kritikk og straff. Den siste påstanden var formulert slik at dersom man var *helt uenig* i at «Sigurd fortjener kritikk for den inntrufne hendelsen», skulle man ha svart at man var *helt enig* i at «det ikke er riktig å kritisere Sigurd i dette tilfellet». For disse tre påstandene (illustrert i tabell under, «*R*» indikerer reversert skala) ble Likert-skalaen reversert til: 1=7, 2=6, 3=5 og 4=4, eksempelvis: «helt uenig» indikerer dermed «helt enig». Alle de avhengige variablene oppnådde høye alfa-koeffisienter, hvilket indikerer at respondentenes svar var konsistente, og målingene var reliable.

Variabel	Cronbach's Alpha (α-verdi)
1. Kritikkverdighet	**,933**
Jeg synes Sigurd fortjener kritikk for den inntrufne hendelsen	
Sigurds fremferd kan og bør kritiseres	
Det er ikke riktig å kritisere Sigurd i dette tilfellet *(R)*	
2. Moralsk ansvar	**,918**
Sigurd er moralsk ansvarlig for feilen	
Det moralske ansvaret for hendelsen må Sigurd ta	
Det blir galt å holde Sigurd moralsk ansvarlig for det som skjedde *(R)*	
3. Straff	**,936**
Sigurd fortjener straff for dette	
Det er riktig at Sigurd utsettes for en disiplinær reaksjon etter dette	
Det blir feil å straffe Sigurd for det som skjedde *(R)*	
4. Sinne	**,930**
Jeg ble sint av å lese om Sigurd	
Det Sigurd gjorde vekket sinne i meg	

Tabell 1. Cronbach's alpha. ((R) indikerer reversert skala)

Korrelasjonsanalyse

Pearons' todimensjonale (bivariat) analyse ble benyttet for å sjekke hvordan ulike variabler korrelerte. Korrelasjonen varierer fra -1 til 1, hvor førstnevnte innebærer en sterk negativ samvariasjon, mens sistnevnte innebærer en sterk positiv samvariasjon. Sammenhengen defineres som svak når tallet er under 0,3, og sterk når tallet er over 0,7 (Larsen, 2017).

Det er lav korrelasjon mellom demografiske data (alder, lønn og utdannelse) og de syv avhengige variablene. Følgelig er det ingen systematisk samvariasjon (i vesentlig grad) mellom de demografiske variablene og deltakernes respons på utfallsmålene (de avhengige variablene). Av plasshensyn ble de demografiske variablene for lederroller, arbeidssektor og kjønn ekskludert da de viste enda lavere korrelasjoner med de avhengige variablene.

Kritikkverdighet, moralsk ansvar og straff har alle sterke positive korrelasjoner. Dette indikerer at jo høyere verdi respondentene ga på en av disse tre avhengige variablene, jo høyere var tendensen til at de også ga en høy verdi på de to andre. Respondenter som mente at Sigurds fremferd var kritikkverdig tenderte også til å plassere moralsk ansvar og ønsket at Sigurd skulle straffes. Dette betyr at jo mer enig respondentene var i en påstand, desto høyere var tendensen til at de var enig i

de to andre påstandene. Den samme tendensen ser man også for opplevd sinne, som korrelerer en del med de overnevnte variablene.

	Variabel	1	2	3	4	5	6	7	8	9
1.	**Kritikkverdighet**	—								
2.	**Moralsk ansvar**	**,89****	—							
3.	**Straff**	**,81****	**,72****	—						
4.	**Sinne**	,52**	,51**	,55**	—					
5.	**Erstatning fra Sigurd**	,28**	,26**	,27**	,27**	—				
6.	**Erstatning fra den norske stat**	-,01	,02	,03	-,01	-,07	—			
7.	**Erstatning fra AI-selskapet**	,01	,04	,13	,1	,18*	-,05	—		
8.	**Alder**	,02	,04	,	,15*	,04	,15*	-,2**	—	
9.	**Årslønn**	-,2**	-,17*	-,18*	-,07	-,11	,13	-,3**	,28**	—
10.	**Utdannelse**	,03	,04	,01	-,03	-,06	,1	-,02	,08	,28**

*Tabell 2. Korrelasjonsmatrise. (*p<,05/**p<,01)*

3.3 Hypotesetesting - test av mellomgruppeforskjeller

Uavhengige t-tester ble gjennomført for å teste de syv hypotesene og avdekke eventuelle mellomgruppeforskjeller.

H1: Respondentene mener at Sigurd i større grad fortjener kritikk når han har et valg (mulighet for manuell overstyring).

En uavhengig t-test ble utført for å undersøke om deltakerne mente at Sigurd i større grad fortjente kritikk når han «hadde et valg» (mulighet for manuell overstyring) enn når han «ikke hadde et valg». Det var signifikant forskjell i resultatene for «hadde et valg» (M=4,295, SD=1,692) og for «hadde ikke et valg» (M=3,768, SD=1,837), forhold; t(203)=2,129, p=.034.

Resultatene gir støtte til hypotesen om at Sigurd i større grad fortjente kritikk når han hadde et valg. Dermed beholdes hypotesen.

H2: Respondentene mener at Sigurd i større grad er moralsk ansvarlig for hendelsen når han har et valg (mulighet for manuell overstyring).

En uavhengig t-test ble utført for å undersøke om deltakerne mente at Sigurd i større grad var moralsk ansvarlig når han «hadde et valg» (mulighet for manuell overstyring) enn når han «ikke hadde et valg». Det var ikke signifikant forskjell i resultatene for «hadde et valg» (M=4,351, SD=1,802) og for «hadde ikke et valg» (M=3,853, SD=1,876), forhold; t(203)=1,930, p=.055.

Resultatene gir ikke støtte til hypotesen om at Sigurd i større grad var moralsk ansvarlig når han hadde et valg, da forskjellen er marginalt for liten til å være statistisk signifikant. Hypotesen forkastes.

H3: Respondentene mener at Sigurd i større grad fortjener straff når han har et valg (mulighet for manuell overstyring).

En uavhengig t-test ble utført for å undersøke om deltakerne mente at Sigurd i større grad fortjente straff når han «hadde et valg» (mulighet for manuell overstyring) enn når han «ikke hadde et valg». Det var ikke signifikant forskjell i resultatene for «hadde et valg» (M=3,333, SD=1,720) og for «hadde ikke et valg» (M=3,070, SD=1,810), forhold; t(203)=1,063, p=.289.

Resultatene støtter ikke hypotesen om at Sigurd i større grad fortjente straff når han hadde et valg. Hypotesen forkastes.

H4: Respondentene føler i større grad på et sinne ovenfor Sigurd når han har et valg (mulighet for manuell overstyring).

En uavhengig t-test ble utført for å undersøke om deltakerne i større grad følte på et sinne når han «hadde et valg» (mulighet for manuell overstyring) enn når han «ikke hadde et valg». Det var ikke signifikant forskjell i resultatene for «hadde et

valg» (M=3,568, SD=1,635) og for «hadde ikke et valg» (M=3,463, SD=1,713), forhold; t(201)=.477, p=.655.

Resultatene støtter ikke hypotesen om at deltakerne i større grad følte et sinne ovenfor Sigurd når han hadde et valg. Hypotesen forkastes.

H5: Respondentene mener at Sigurd i større grad bør pålegges et erstatningsansvar når han har et valg (mulighet for manuell overstyring).

En uavhengig t-test ble utført for å undersøke om deltakerne i større grad mener at Sigurd bør pålegges et erstatningsansvar når han «hadde et valg» (mulighet for manuell overstyring) enn når han «ikke hadde et valg». Det var ikke signifikant forskjell i resultatene for «hadde et valg» (M=3,59, SD=2,262) og for «hadde ikke et valg» (M=3,05, SD=2,181), forhold; t(201)=1,740, p=.083.

Resultatene gir ikke støtte til hypotesen om at Sigurd i større grad bør pålegges et erstatningsansvar når han hadde et valg. Det er en viss forskjell, men heller ikke denne er signifikant nok. Hypotesen forkastes.

H6: Respondentene mener at den norske stat i mindre grad bør pålegges et erstatningsansvar når Sigurd har et valg (mulighet for manuell overstyring).

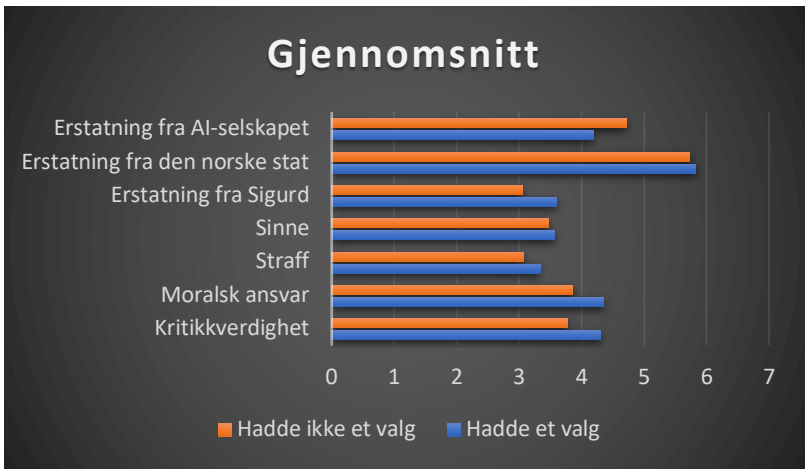
En uavhengig t-test ble utført for å undersøke om deltakerne mener at den norske stat i mindre grad bør pålegges et erstatningsansvar når Sigurd «hadde et valg» (mulighet for manuell overstyring) enn når han «ikke hadde et valg». Det var ikke signifikant forskjell i resultatene for «hadde et valg» (M=5,82, SD=1,337) og for «hadde ikke et valg» (M=5,73, SD=1,508), forhold; t(201)=.445, p=.657.

Resultatene gir ikke støtte til hypotesen om at den norske stat i mindre grad bør pålegges et erstatningsansvar når Sigurd hadde et valg. Her tenderer ikke resultatene i retning av hypotesen, men heller i motsatt retning, dog er forskjellene omtrent ubetydelige. Hypotesen forkastes.

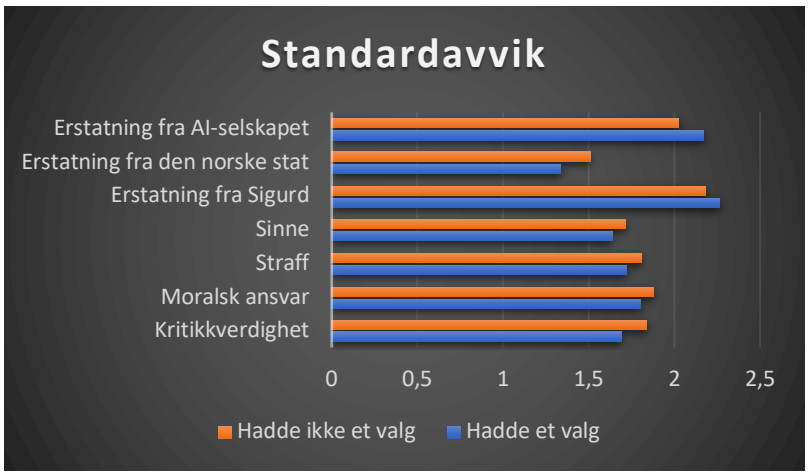
H7: Respondentene mener at selskapet som lager den kunstige intelligente algoritmen i mindre grad bør pålegges et erstatningsansvar når Sigurd har et valg (mulighet for manuell overstyring).

En uavhengig t-test ble utført for å undersøke om deltakerne mener at selskapet som lager den kunstig intelligente algoritmen i mindre grad bør pålegges et erstatningsansvar når Sigurd «hadde et valg» (mulighet for manuell overstyring) enn når han «ikke hadde et valg». Det var ikke signifikant forskjell i resultatene for «hadde et valg» (M=4,19, SD=2,168) og for «hadde ikke et valg» (M=4,71, SD=2,024), forhold; t(203)=-1,771, p=.078.

Resultatene gir ikke støtte til hypotesen om at selskapet som lager den kunstig intelligente algoritmen i mindre grad bør pålegges et erstatningsansvar når Sigurd hadde et valg. Forskjellen er tilstedeværende, men ikke signifikant nok. Hypotesen forkastes.



Figur 7. Gjennomsnitt i utvalgene.



Figur 8. Standardavvik i utvalgene.

4.0 Diskusjon

Målet med studien har vært å belyse hvor ansvaret plasseres når en kunstig intelligent militær-drone gjør fatal feil. Funnene viser at tilstedeværelse av valgmuligheter ikke spiller en vesentlig rolle for menneskers ansvarsfordeling i denne studien. På bakgrunn av studiens resultater vil det diskuteres hvorfor hovedfunnene avviker fra tidligere forskning på ansvarsfordeling, hvor «valgfenomenet» spiller en betydelig rolle, samt belyses hvor ansvaret i denne studien plasseres.

4.1 Oppsummering av hovedfunn

Hypotese 1 handlet om i hvilken grad respondentene mente at Sigurd fortjente kritikk for den inntrufne hendelsen. Resultatene tilsa at respondentene mente at Sigurd i større grad fortjente kritikk når han hadde et valg. Hypotesen beholdes.

Hypotese 2 tok for seg i hvilken grad respondentene mente Sigurd var moralsk ansvarlig for den inntrufne hendelsen. Resultatene tilsa at respondentene ikke mente at Sigurd i større grad var moralsk ansvarlig når han hadde et valg. Forskjellen var marginalt for liten til å være statistisk signifikant, og resultatene tenderte i retning av hypotesen. Hypotesen forkastes.

Hypotese 3 handlet om i hvilken grad respondentene mente at Sigurd fortjente straff for den inntrufne hendelsen. Resultatene tilsa at respondentene ikke mente at Sigurd i større grad fortjente straff når han hadde et valg. Mellomgruppeforskjell var liten, men tenderte i retning av hypotesen. Hypotesen forkastes.

Hypotese 4 dreide seg om i hvilken grad respondentene følte et sinne overfor Sigurd. Resultatene tilsa at respondentene ikke følte en større grad av sinne overfor Sigurd når han hadde et valg. Mellomgruppeforskjell var liten, men tenderer i retning av hypotesen. Hypotesen forkastes.

Hypotese 5 handlet om i hvilken grad respondentene mente at Sigurd bør pålegges et erstatningsansvar. Resultatene tilsa at respondentene ikke mente at Sigurd i større grad bør pålegges et erstatningsansvar når han hadde et valg.

Mellomgruppeforskjellen var marginalt for liten til at hypotesen kunne beholdes, og resultatene tenderte i retning av hypotesen. Hypotesen forkastes.

Hypotese 6 tok for seg i hvilken grad respondentene mente at den norske stat bør pålegges et erstatningsansvar. Resultatene tilsa at respondentene ikke mente at den norske stat i mindre grad bør pålegges et erstatningsansvar når Sigurd hadde et valg. For denne påstanden var mellomgruppeforskjellene lavest, og nesten fraværende. Resultatene tenderte ikke i retning av hypotesen, men heller i motsatt retning, dog er forskjellene omtrent ubetydelige. Det er verdt å merke seg at gjennomsnittet var høyt i begge utvalgene, og standardavvikene relativt lave, sammenlignet med de andre hypotesene. Dette indikerer at respondentene mente at den norske stat skal holdes erstatningsansvarlig. Hypotesen forkastes.

Hypotese 7 dreide seg om i hvilken grad respondentene mente at selskapet som lager den kunstige intelligens bør pålegges et erstatningsansvar. Resultatene tilsa at respondentene ikke mente at selskapet i mindre grad bør pålegges et erstatningsansvar når Sigurd hadde et valg. Mellomgruppeforskjellen var marginalt for liten til at hypotesen kunne beholdes, og resultatene tenderte i retning av hypotesen. Hypotesen forkastes.

H1 beholdes, mens de andre hypotesene forkastes, da mellomgruppeforskjellene ikke var statistisk signifikante nok. Det er likevel verdt å nevne at det for alle hypotesene var visse mellomgruppeforskjeller, hvorav samtlige tenderer i retningen av hypotesene, med unntak av H6. I tillegg er det verdt å merke seg at H2, H5 og H7 var marginalt for lite signifikante til å kunne beholdes, med p-verdier på henholdsvis 0,055; 0,083; 0,078.

4.2 Teoretiske implikasjoner

I vignettundersøkelsen fremgår det tydelig at AI-en, som tar avfyringsbeslutninger, er langt mer treffsikker (gjør mindre feil) og effektiv enn mennesker. Det kan derfor i vesentlig grad argumenteres for at Sigurd, i tilfellet hvor han hadde mulighet til manuell overstyring, gjorde rett i å ikke endre innstillingene, og la AI-en ta beslutninger.

Vallentyne (2008) sin andre betingelse sier at man ikke er personlig ansvarlig, dersom man bare kunne avverget utfallet ved en urimelig stor ulempe. I Cappelen

et al., (2016) sin studie fremstilles denne betingelsen som et «forced choice», hvor det trygge alternativet fremstår mye dårligere (25 NOK utbetalt mot lotteriet som har en forventningsverdi på 400 NOK). Valget til Sigurd kan også sies å være fremtvunget ettersom manuell overstyring beviselig fremsto som et mye dårligere alternativ enn å la den sofistikerte AI-en ta beslutninger. Sigurd var fullt klar over at å la AI-en ta beslutninger medførte vesentlig høyere sannsynlighet for riktige beslutninger. Således harmonerer dette med Cappelen et al., (2016) sin «forced choice treatment». Hvorvidt manuell overstyring er et urimelig alternativ kan likevel diskuteres. Gitt den sofistikerte kunstige intelligensens påviselige overlegenhet hadde det vært umoralsk av Sigurd å ikke benytte seg av den, fordi sannsynligheten for galt utfall øker. Hvilket Dyndal et al., (2017) også problematiserer. De mente at det kan være moralsk gunstig å bruke autonom droneteknologi ettersom de kan prosessere vesentlig mer informasjon enn mennesker, og derfor ta velbegrunnede beslutninger. Droner påvirkes heller ikke av støy eller følelser, hvilket kan redusere risikoen for krigsforbrytelser.

Vallentyne (2008) sin første betingelse tilsier at man ikke skal holdes personlig ansvarlig dersom man ikke kunne endret sannsynligheten for utfallet ved å velge annerledes. Dersom Sigurd hadde valgt manuell overstyring, ville sannsynligheten for et galt utfall økt, nettopp fordi AI-en hadde langt bedre forutsetninger for å ta riktige beslutninger. Derfor kan det hevdes at han heller ikke kunne endret sannsynlighet for utfallet. Rent hypotetisk kunne Sigurd selvfølgelig endret sannsynligheten for utfallet ved å ikke utføre jobben sin, men her legges det til grunn at han hadde to valg; manuell overstyring eller la AI-en ta beslutninger selv.

Vallentyne (2008) sine betingelser fanger opp situasjoner hvor individer på forhånd ikke har noen grunn til å velge annerledes. Det samme må kunne sies å gjelde for Sigurd, som heller ikke hadde noen grunn til å velge annerledes. Spørsmålet blir derfor hvor fritt og selvstendig valget til Sigurd i utgangspunktet var, eller, om han i realiteten hadde et valg. Vallentyne (2008) hevder at individer som tar autonome valg, og har full kunnskap om konsekvensene, skal stå ansvarlig for utfallet. Var respondentens oppfatning i denne studien at Sigurd sitt valg var autonomt? Hvis ikke, kan det ha vært en årsak til at det gjennomgående i studien var relativt lav plassering av ansvar hos Sigurd for begge utvalgene?

Sigurd var ansatt i Forsvaret og hans mandat, som underordnet, var å bekjempe terrorisme. Det er ikke urimelig å anta at respondentene var av den oppfatning at

tilværelsen som soldat generelt karakteriseres ved lite valgfrihet og stor grad av ordreadlydning. Hvilket også implisitt kommer til utrykk i resultatdelen. Det fremgår at gjennomsnittet lå høyest ved plassering av erstatningsansvar hos den norske stat for begge utvalgene. For dette utfallsmålet var det også lavest mellomgruppeforskjell, og lavest standardavvik for begge utvalgene (se figur 7 og 8). For utvalget som ble forespeilet at Sigurd hadde et valg lå gjennomsnittet på 5,82, mens gjennomsnittet for utvalget som ble forespeilet at Sigurd ikke hadde et valg var 5,73. Dette betyr at den «gjennomsnittlige respondent» – uavhengig av om Sigurd hadde valgmuligheter eller ikke – mener at den norske staten skal stå erstatningsansvarlig for utfallet. De nokså lave standardavvikene, relativt til standardavvikene for de andre hypotesene, indikerer også en samstemt oppfatning av dette. Hva som ligger bakenfor respondentenes tilbøyelighet for plassering av erstatningsansvar hos den norske stat, kan være mye. Likevel, det er ikke utenkelig at de nettopp er av den oppfatning at Sigurd sitt valg var heteronomt, og at man som soldat er en del av et større system, hvor det kontinuerlig løper en kommandolinje – dermed skal heller ikke Sigurd stå ansvarlig for utfallet. Fischer & Ravizza (1998) presenterte det de kalte for «excusing conditions» i tilknytning til moralsk ansvar. Et slikt forhold handler om tvang. Selv om Sigurd fritt har valgt jobben som soldat, er det grunn til å hevde at både valget om å la AI-en ta beslutninger eller ikke, og arbeidet som soldat generelt, er noe Sigurd er pålagt av den norske stat, og således også «tvunget» til.

En distinkt forskjell mellom denne studien og Cappelen et al., (2016) sin studie, er det faktum at Sigurd sitt valg ikke påvirket ham direkte. Eller sagt på en annen måte: Sigurd handlet på vegne av en tredjepart, Forsvaret og den norske stat, mens deltakerne i studien til Cappelen et al., (2016) handlet på vegne av seg selv, og deres valg påvirket heller ingen andre. At Sigurd handlet på vegne av den norske stat må kunne forstås som en medvirkende faktor til at respondentene i vesentlighet påla den norske stat et erstatningsansvar. I tillegg var deltakerne i studien til Cappelen et al., (2016) kausalt ansvarlig for utfallene, i motsetning til i denne studien, hvor det var dronen som utgjorde den avgjørende årsaken til utfallet. Vallentyne (2008) problematiserer i sin studie situasjoner hvor det oppstår delansvar, det vil si, flere parter har sine årsaksbidrag til det endelige utfallet. Dette manifesterte seg også i denne studien, hvor Sigurd sitt valg bidro til det endelige utfallet. Det var likevel ikke valget til Sigurd som var den direkte

årsaken til utfallet, da det var AI-en, som ved en feiltakelse, tok en gal beslutning. Vallentyne (2008) hevder at individer bare er delvis ansvarlig når deres valg fører til et skifte på mindre enn 100 prosent i sannsynligheten for det aktuelle utfallet. Vallentyne (2008) sin tilnærming om delansvar, og det faktum at Sigurd ikke var kausalt ansvarlig for utfallet alene, kan ha vært en medvirkende årsak til at mellomgruppeforskjellene ikke var signifikante i tildeling av ansvar overfor Sigurd.

Anvendelse av autonom AI-teknologi skaper en utydeliggjøring av hvor ansvaret egentlig skal plasseres. Slike situasjoner karakteriseres som «moralske ansvarshull» fordi det blir mindre tydelig hvordan ansvaret skal fordeles (Matthias, 2003; Dyndal et al., 2017). Tradisjonelt holdes produsenter og/eller operatører moralsk og legalt ansvarlig for feil begått av maskiner (Matthias, 2003). Situasjonen er derimot annerledes i denne undersøkelsen fordi dronen baserer seg på algoritmiske beslutninger, som verken AI-selskapet, Forsvaret eller Sigurd i prinsippet var i stand til å forutse. Selv om AI-en var mer sofistikert enn mennesker, var den ikke ufeilbarlig. Siden det ikke er mulig å plassere et moralsk ansvar hos AI-en, mente respondentene at (erstatnings)ansvaret skal plasseres hos de som har valgt å ta i bruk teknologien, altså den norske stat. Det kan derfor trekkes en parallell til det filosofen Peter French hevdet, nemlig at organisasjoner absorberer og utstråler menneskers intensjoner og handlinger, og at beslutningene derfor er organisatoriske. Organisasjoner har et moralsk ansvar uavhengig av om intensjonen reflekterer selve handlingen (Shaw & Barry, 2015, s.207).

Akkurat som i Dietvorst et al., (2014) sin studie på «algoritme-aversjon», som viste at mennesker systematisk favoriserer menneskelige beslutningstakere, ble respondentene i denne studien også forespeilet en algoritme som gjør feil, men i utgangspunktet tar flere riktige beslutninger enn mennesker. Selv om de ble forespeilet denne situasjonen ansvarliggjorde de ikke Sigurd mer når han kunne velge å avstå fra å benytte seg av algoritmen. Det laveste gjennomsnittet for begge utvalgene var for utfallsmålet som gjald om Sigurd fortjente straff for den inntrufne hendelsen. Det norske lovverket er relativt tydelig på at bevisste handlinger straffes hardere enn utilsiktede gjerninger. Som Gray, Young & Watz (2012) presiserte, betyr intensjoner knyttet til handlinger mye for skyld- og ansvarsspørsmål. Hvilket Woolfolk, Doris & Darley (2006) fant i sitt eksperiment, hvor mennesker som tvinges til å henrette andre, men også ønsker dem døde,

oppfattes som mer umoralske. Det er liten tvil om at intensjonene til Sigurd ikke var onde, noe som antagelig er vesentlig for forståelsen av at respondentene ikke ønsket å straffe Sigurd, uavhengig av tilstedeværelse av valgmuligheter eller ikke.

I Cappelen et al., (2016) sin studie fremgikk det at tredjepartene tilla en normativ betydning av at deltakerne hadde tatt et valg. I sum antyder fravær av mellomgruppeforskjeller at respondentene i denne studien ikke tillegger en normativ betydning av at Sigurd kunne ta valg/tok et valg.

4.2 Implikasjoner for stater/organisasjoner/selskaper

Funnene i studien viser at den norske stat står erstatningsansvarlig for fatale feil, som en sofistikert og autonom kunstig intelligens forårsaker. Gitt at teknologien er mer sofistikert enn menneskelig ekspertise, spiller det heller ingen vesentlig rolle for ansvarsfordelingen om operatøren kan velge å benytte seg av teknologien, eller ikke. I bredere sammenheng impliserer funnene at ansvaret faller på staten/organisasjonen/selskapet ved implementering av autonome og algoritmisk beslutningssystemer. Det er derfor viktig at de ovenfornevnte er klar over at implementering av kunstig intelligent og autonom beslutningsteknologi, som kan volde sivil skade, medfører et betydelig ansvar, og kanskje enda større enn ved tradisjonell beslutningstaking, som i stor grad beror på operatøren sine evner.

4.3 Metodiske begrensninger og kritikk av studien

Eksperimentet var av hypotetisk karakter og bygget på en fiktiv fortelling som daterer seg frem i tid – 2025. Respondentene besvarte således vignettundersøkelsen med dette «in mente» og det er ikke utenkelig at enkelte deltakere synes situasjonen kan ha vært vanskelig å se for seg, eller tenke direkte konsekvenser av. Undersøkelsen ble ikke gjennomført under kontrollerte forhold, hvilket eksempelvis kan ha preget deltakernes refleksjoner over de ulike påstandene og dermed også svarene, og resultatene. Ettersom undersøkelsen var anonym kan det heller ikke utelukkes at respondenter kan ha gjennomført undersøkelsen mer enn én gang. Som det fremgikk i deskriptive data var det en overvekt av respondenter med høyere utdanning, hvilket ikke harmonerer med populasjonen som helhet.

Utviklingen av kunstig intelligens og autonome systemer innenfor militære våpensystemer går raskere enn forskningen. Derfor har det også vært krevende å finne relevant empirisk materiale. Det er også relativt lite tilgjengelig informasjon om FoU på dette området ettersom dette foregår, naturlig nok, innenfor nokså lukkede systemer.

4.4 Anbefalinger til videre forskning

I videre forskning kan det være formålstjenlig å teste problemstillingen på et større utvalg for å sikre en bredere forståelse av menneskers ansvarsfordeling. Det kunne også vært interessant å teste problemstillingen på tvers av land for å avdekke om «valgfenomenet» har ulik betydning for ansvarsfordeling. I tillegg kunne det vært interessant å inkludere en kontrollgruppe som blir forespeilt den samme historien, men hvor kunstig intelligens ekskluderes, for å se hvordan kontrollgruppen fordeler ansvar når en menneskelig agent alene står kausalt ansvarlig for det samme utfallet. Særlig fordi begge utvalgene i denne studien holdt den norske stat erstatningsansvarlig, ville det vært interessant å se om resultatene i en slik kontrollgruppe avviker vesentlig fra disse.

MIT har utviklet en plattform kalt «Moral Machine» for å teste hvordan menneskers stiller seg til ulike moralske dilemmaer tilknyttet AI, som selvkjørende biler. Det ville vært interessant å teste studiens problemstillingen med et forskningsdesign, som ligner dette, med mer virkelighetsnære og realistiske simuleringer eller videofremstillinger.

Utfallet i dette eksperimentet var fatalt og det kan ikke utelukkes at det hadde vært større mellomgruppeforskjeller dersom konsekvensene var mindre. Det kunne vært nyttig å teste samme problemstilling, men med et mindre alvorlig utfall.

5.0 Konklusjon

Hovedformålet med denne studien har vært å besvare det overordnede forskningsspørsmålet:

«Hvor plasseres ansvaret når kunstig intelligent militær-drone gjør fatal feil?»

Det ble utviklet syv hypoteser, som et bidrag til å besvare denne problemstillingen. Hypotesene ble utviklet i tro på at tilstedeværelse av valgmuligheter i vesentlig grad spiller en viktig rolle i menneskers ansvarsplassering, slik det fremgikk i studiene til Cappelen et al., (2016) og Savani & Rattan (2012). Resultatene viser at dette ikke stemte for denne studien. Til tross for at deltakerne stilte seg mer kritisk til Sigurd når han hadde et valg, var det ikke signifikante mellomgruppeforskjeller som impliserte at deltakerne i de to utvalgene hadde vesentlig forskjellige oppfatninger og holdninger til Sigurd sin fremferd. Resultatene i undersøkelsen tenderte likevel i retning av hypotesene, det samme indikerte sammenhengen mellom de avhengige variablene (kritikkverdighet, moralsk ansvar og straff), som viste relativt høye positive korrelasjoner.

Funnene i studien er entydige på at den norske stat i vesentlig grad pålegges et erstatningsansvar for den fatale feilen. I bredere sammenheng impliserer dette at stater/organisasjoner/selskaper bør være klar over at implementering av kunstig intelligent og autonom beslutningsteknologi, kan medføre en betydelig ansvarliggjøring når disse gjør fatale feil – kanskje i enda større grad enn ved tradisjonell beslutningstaking, som i stor grad beror på operatøren sine evner.

6.0 Litteraturliste

Almås, G.B. (2019, 5. februar). Digitalt diktatur: Kina planlegger sosialt poengsystem. *Norges rikskringkasting.* Hentet fra: https://www.nrk.no/urix/kinas-digitale-diktatur_-gar-du-pa-rodt-lys_-blir-du-uthengt-pa-storskjerm-1.14369439

Andersen, L. M. & Bakkeli, M. (2015). Big Data: Hva er Big Data, og hva betyr Big Data for deg?. *PWC Consulting*. Hentet fra: https://www.pwc.no/no/publikasjoner/information-management/big-data.pdf

Arstad, S. (2019, 1. mars). Norsk drone-eventyr. *Forsvarets Forum.* Hentet fra: https://forsvaretsforum.no/nyhetsartikkel/norsk-drone-eventyr/104377

Bathaee, Y. (2018). The Artificial Intelligence Black Box and the Failure of Intent and Causation. *Harvard Journal of Law & Technology. 31*, 50. Hentet fra: https://jolt.law.harvard.edu/assets/articlePDFs/v31/The-Artificial-Intelligence-Black-Box-and-the-Failure-of-Intent-and-Causation-Yavar-Bathaee.pdf

Brownell, K.D., Kersh, R., Ludwig, D.S., Post, R.C., Puhl, R.M., Schwartz, M.B. & Willett, W.C. (2010). Personal responsibility and obesity: A constructive approach to a controversial issue. *Health Affairs*, 29(3): 379–387. DOI: 10.1377/hlthaff.2009.0739

Cappelen, A. W., Fest, S., Tungodden, B., & Sørensen, E. Ø. (2016) Choice and personal responsibility: What is a morally relevant choice? *NHH Dept. of Economics Discussion paper*, 27, 2014

Chui, M., Manyika, J., & Miremadi, M. (2017, 12. april). The Countries Most (and Least) Likely to be Affected by Automation. *Harvard Business Review*. Hentet fra: https://hbr.org/2017/04/the-countries-most-and-least-likely-to-be-affected-by-automation

Coglianese, C. & Lehr, D. (2016). Regulating by robot: Administrative Decision Making in the Machine-Learning Era. *Faculty Scholarship at Penn Law*, 1147-1223. Hentet fra:

https://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=2736&context=faculty_scholarship

Confessore, N. (2018, 4.april). Cambridge Analytica and Facebook: The Scandal and the Fallout So Far. *The New York Times.* Hentet fra: https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html?fbclid=IwAR3wYwb7fcFm5_Ji1vUNRlq2NaKHgvsFyIz7Ggs80G-COHCgXA3jH7rhPNI

Cuthberson, A. (2019, 6. november). Self-driving Uber in fatal crash did not know people could jaywalk. *Independent*. Hentet fra: https://www.independent.co.uk/life-style/gadgets-and-tech/news/uber-self-driving-crash-death-elaine-herzberg-arizona-jaywalk-a9187791.html?fbclid=IwAR18utMp5XnoZ7ub_AHKWfbFFI_bKl1YqC5kfzQ9dvRNFJjsfEfjWtjcXyU

Davenport, T. H. & Harris, J. G. (2007*). Competing on Analytics: The New Science of Winning* (1st Edition). Massachusetts: Harvard Business School Press

Dawes, R.M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7), 571-582. https://doi.org/10.1037/0003-066X.34.7.571

Dawes, R.M., Faust, D., & Meehl, P.E. (1989). Clinical versus actuarial judgment. *Science*, 243(4899), 1668–1674. DOI: 10.1126/science.2648573

De Nasjonale Forskningsetiske Komiteene. (2006). Forskingsetiske Retningslinjer for Samfunnsvitenskap, Humanoria, Juss og Teologi. *De Nasjonale Forskningsetiske Komiteene; Forskningsetiske Retningslinjer (Mars).*

Dietvorst, B.J., Simmons, J., & Massey, C. (2014). Understanding algorithm aversion: Forecasters erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General,* 144(1), 114.

Dietvorst, B.J., Simmons, J.P., & Massey, C. (2016). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3), 1155–1170.

Donnelly, L. (2017, 7. mars). ‘Forget your GP, robots will 'soon be able to diagnose more accurately than almost any doctor’. *The Telegraph.* Hentet fra: https://www.telegraph.co.uk/technology/2017/03/07/robots-will-soon-able-diagnose-accurately-almost-doctor/

Doyle, S. (2018, 6. november). Drone warfare: the autonomous debate. *Engineering & Technology*. Hentet fra: https://eandt.theiet.org/content/articles/2018/11/drone-warfare-the-autonomous-debate/

Drone Wars. (2012). UK Drone Strike Stats. *Drone Wars*. Hentet fra: https://dronewars.net/uk-drone-strike-list-2/

Dyndal, G. L., Berntsen, T.A. & Redse-Johansen, S. (2017, 27. juli). Autonomous military drones: no longer science fiction. *NATO Review.* Hentet fra: https://www.nato.int/docu/review/articles/2017/07/28/autonomous-military-drones-no-longer-science-fiction/index.html

Einhorn, H.J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance*, 7(1), 86–106. https://doi.org/10.1016/0030-5073(72)90009-8

European Commision. (2019). A definition of Artificial Intelligence: Main Capabilities and Scientific Disciplines. Hentet fra: https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Gibbs, S. (2015, 27. juli). Musk, Wozniak and Hawking urge ban on warfare AI and autonomous weapons. *The Guardian.* Hentet fra: https://www.theguardian.com/technology/2015/jul/27/musk-wozniak-hawking-ban-ai-autonomous-weapons

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological inquiry*, 23(2), 101-124. https://doi.org/10.1080/1047840X.2012.651387

Griffin, A. (2015, 12. januar). Stephen Hawking, Elon Musk and others call for research to avoid dangers of artificial intelligence. *Independent.* Hentet fra:

https://www.independent.co.uk/life-style/gadgets-and-tech/news/stephen-hawking-elon-musk-and-others-call-for-research-to-avoid-dangers-of-artificial-intelligence-9972660.html

Gripsrud, G., Olsson, U., & Silkoset, R. (2010). Metode og dataanalyse : Beslutningsstøtte for bedrifter ved bruk av JMP (2. utg.). Kristiansand: Høyskoleforl.

Gripsrud, G., Olsson, U., & Silkoset, R. (2016). *Metode og dataanalyse : Beslutningsstøtte for bedrifter ved bruk av JMP, Excel og SPSS* (3. utg.). Oslo: Cappelen Damm akademisk.

Hawking, S., Russel, S., Tegmark, M. & Wilczek, F. (2014, 1. mai). Stephen Hawking: 'Transcendence looks at the implications of artificial intelligence - but are we taking AI seriously enough?'. *Independent.* Hentet fra: https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html

Hexmoor, H. (2013). *Essential Principles for Autonomous Robotics.* California: Morgan & Claypool Publishers

Hofstadter, D.R. (1980). *Gödel, Escher, Bach: An eternal golden braid* (1st Vintage Books ed.). New York: Vintage Books.

Jensen, S. (2016, april). *Den fjerde industrielle revolusjon – muligheter til å bedre ressursutnyttelsen*. Innlegg presentert ved IKT-Norges konferanse, Oslo. Sammendrag hentet fra: https://www.regjeringen.no/no/aktuelt/den-fjerde-industrielle-revolusjon--muligheter-til-a-bedre-ressursutnyttelsen/id2483283/

Kommunal- og moderniseringsdepartementet. (2020). *Nasjonal strategi for kunstig intelligens.* Hentet fra: https://www.regjeringen.no/contentassets/1febbbb2c4fd4b7d92c67ddd353b6ae8/no/pdfs/ki-strategi.pdf

Larsen, A. (2017). *En enklere metode : Veiledning i samfunnsvitenskapelig forskningsmetode* (2. utg.). Bergen: Fagbokforlaget

Lohr, S. (2016, 17. oktober). IBM Is Counting on Its Bet on Watson, and Paying Big Money for It. *The New York Times*. Hentet fra: https://www.nytimes.com/2016/10/17/technology/ibm-is-counting-on-its-bet-on-watson-and-paying-big-money-for-it.html

Matthias, A. (2003). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183. Hentet fra: https://www.academia.edu/243900/The_Responsibility_Gap._Ascribing_Responsibility_for_the_Actions_of_Learning_Automata

McAfee, A. & Brynjolfsson, E. (2012, oktober). Big Data: The Management Revolution. *Harvard Business Review.* Hentet fra: https://hbr.org/2012/10/big-data-the-management-revolution

Meehl, P.E. (1954). *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.* Minnesota: University of Minnesota Press

Moffitt, R. A. (2015). The deserving poor, the family, and the U.S. welfare system. *Demography*, 52(3): 729–749. DOI: 10.1007/s13524-015-0395-0

NOU, 2016:3. (2016). *Ved et vendepunkt: Fra ressursøkonomi til kunnskapsøkonomi.* Hentet fra: https://www.regjeringen.no/contentassets/64bcb23719654abea6bf47c56d89bad5/no/pdfs/nou201620160003000dddpdfs.pdf

Rao, A,S. & Verweij, G. (2017). Sizing the prize: What's the real value of AI for your business and how can you capitalize?. *PWC.* Hentet fra: https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf

Read, R. (2020, 3. januar). World's most feared drone: CIA's MQ-9 Reaper killed Soleimani. *Washington Examiner*. Hentet fra: https://www.washingtonexaminer.com/policy/defense-national-security/worlds-most-feared-drone-cias-mq-9-reaper-killed-soleimani?fbclid=IwAR2GtTkkskENeHRnYa7DlH409IqdVcynz5hlz2Oe-3tbTYpX_-h6-ZdGhds

Russel, S., Dewey, D. & Tegmark, M. (2015). Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine* 36(4), 105-114. https://doi.org/10.1609/aimag.v36i4.2577

Sabbagh, D. (2019, 18. november). Killer drones: how many are there and who do they target?. *The Guardian*. Hentet fra: https://www.theguardian.com/news/2019/nov/18/killer-drones-how-many-uav-predator-reaper

Savani, K. & Rattan, A. (2012). A choice mind-set increases the acceptance and maintenance of wealth inequality. *Psychological Science*, 23(7): 796–804. https://doi.org/10.1177/0956797611434540

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin*, 66(3), 178-200. https://doi.org/10.1037/h0023624

Scharre, P. (2018). *Army of none: Autonomous Weapons and the Future of War* (1st Edition). New York: W. W. Norton & Company

Scharre, P. (2019, 16. april). Killer Apps: The Real Dangers of an AI Arms Race. *Foreign Affairs.* Hentet fra: https://www.foreignaffairs.com/articles/2019-04-16/killer-apps

Shaw, W. H., & Barry, V. (2015). *Moral issues in business* (13th Edition). Boston: Cengage Learning.

Statistisk sentralbyrå. (2019). *Befolkningens utdanningsnivå.* Hentet fra: https://www.ssb.no/utniv/

Statistisk sentralbyrå. (2019). 11536: *Årslønn, etter sektor, statistikkvariabel og år.* Hentet fra: https://www.ssb.no/statbank/table/11536

Statistisk sentralbyrå. (2020). 12907*: Sysselsatte per 4. kvartal, etter sektor, kjønn, statistikkvariabel og år.* Hentet fra: https://www.ssb.no/statbank/table/12907/

Temming, M. (2020, 14. februar). AI can predict which criminals may break laws again better than humans. *Science News.* Hentet fra: https://www.sciencenews.org/article/ai-can-predict-criminals-repeat-offenders-better-than-humans

Tolfsen, C. (2018, 7. Februar). – Robotene kan bli mer moralske enn oss. *Norges Rikskringkasting.* Hentet fra: https://www.nrk.no/kultur/xl/nar-dataprogrammene-far-moral-1.13867044

Turing, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. Mind, LIX (236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

UN. (2019, 25.mars). Autonomus weapons that kills must be banned, insist UN chief. *UN News*. Hentet fra: https://news.un.org/en/story/2019/03/1035381

Vallentyne, Peter (2008). Brute luck and responsibility. *Politics, Philosophy & Economics*, 7(1): 57–80. https://doi.org/10.1177/1470594X07085151

Wikler, D. (2002). Personal and social responsibility for health. *Ethics and International Affairs*, 16(2): 47–55. https://doi.org/10.1111/j.1747-7093.2002.tb00396.x

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*, 100(2), 283-301. https://doi.org/10.1016/j.cognition.2005.05.002

World Economic Forum. (2017). *The Global Risks Report 2017* (12th Edition). Hentet fra: https://www.deslibris.ca/ID/10090180

Aase, K.A. (2020, 24.februar). Kunstig intelligens varslet om det nye coronaviruset. *Verdens Gang.* Hentet fra: https://www.vg.no/spesial/c/stories/8mKOKE