Handelshøyskolen BI

BTH 36201 Bacheloroppgave - Økonomi og administrasjon

Bachelor thesis 100% - B

Predefinert informasjon

Startdato:	09-01-2023 09:00 CET	**Termin:**	202310
Sluttdato:	01-06-2023 12:00 CEST	**Vurderingsform:**	Norsk 6-trinns skala (A-F)
Eksamensform:	D		
Flowkode:	202310||10737||IN17||B||D		
Intern sensor:	(Anonymisert)		

Navn:

Oda Elin Krag-Stubberud

Informasjon fra deltaker

Tittel *:	Tillit til kunstig intelligens i utdanningssektoren
Navn på veileder *:	Tarjei Alvær Heggernes

Inneholder besvarelsen konfidensielt materiale?:	Nei	**Kan besvarelsen offentliggjøres?:**	Ja

Gruppe

Gruppenavn:	(Anonymisert)
Gruppenummer:	107
Andre medlemmer i gruppen:	Deltakeren har innlevert i en enkeltmannsgruppe

Bacheloroppgave

ved Handelshøyskolen BI

Bilde hentet fra https://questmite.com/technology/grappling-with-the-prospect-of-ai-in-education/

- Tillit til kunstig intelligens i utdanningssektoren -

Eksamenskode og navn:

BTH3620 – Bacheloroppgave – Økonomi og administrasjon

Utleveringsdato:

09.01.2023

Innleveringsdato:

01.06.2023

Studiested:

BI Bergen

«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

I denne bacheloroppgaven undersøkes studenters tillit til å benytte kunstig intelligens innenfor utdanning. Formålet med oppgaven har vært å besvare følgende problemstilling:

I hvilken grad har studenter tillit til at en kunstig intelligens kan være rettferdig i en evalueringsprosess?

Besvarelsen er fundamentalt konstruert etter teori og tidligere forskning som er redegjort i oppgavens teoretiske del. Dette gir en solid teoretisk forankring i studien, og sikrer empiri. Sentrale begreper og delproblemer fra oppgavens teoretiske del dannet grunnlaget for å besvare problemstillingen. Basert på dette ble det identifisert flere variabler som kan påvirke graden av tillit blant studenter. For å undersøke variablene ble det formulert 9 hypoteser. Hypotesene ble undersøkt gjennom ulike statistiske analyser, inkludert en regresjonsanalyse, t-tester, krysstabulering og kji-kvadrattest.

For å belyse den aktuelle problemstillingen og undersøke hypotesene er det innhentet både kvantitativ og kvalitativ data. Det er gjennomført en kvantitativ spørreundersøkelse, hvor 302 respondenter deltok. I tillegg er det innhentet kvalitativ data gjennom fire dybdeintervjuer. Resultater og funn fra den innsamlede dataen danner grunnlaget for å undersøke hypotesene og besvare problemene.

Resultater og funn som er avdekket i forskningen, ga grunnlag for å konkludere med at studenter har en moderat tillit til at kunstig intelligens kan være rettferdig i en evalueringsprosess. I tillegg er det identifisert flere hindringer som gjør det utfordrende å oppnå tillit til systemet, som bias, black box-problemet og algoritme aversjon. Videre indentifiseres funn som indikerer at studenters tillit kan økes ytterliggere dersom det iverksettes tiltak for å redusere bias og algoritme aversjon, og å løse black box-problemet.

Forord

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

Først og fremst vil jeg takke min veileder Tarjei Alvær Heggernes, som har vært til stor hjelp i arbeidet med denne bacheloroppgaven. Videre vil jeg takke alle som besvarte spørreundersøkelsen, og deltok i dybdeintervjuene. Den omfattende dybden i forskningen ville ikke vært mulig uten deres bidrag og engasjement.

Temaet for oppgaven vekket sterkt engasjement i meg, og det har vært utrolig spennende å få muligheten til å gå i dybden på temaet. Det har vært både utfordrende og tidkrevende å skrive en så stor oppgave alene, men prosessen har likevel vært veldig givende og lærerik. Jeg har lært mye om både meg selv og temaet. Å skrive bacheloroppgaven alene har gitt meg verdifull erfaring som jeg ikke ville vært foruten.

Jeg er utrolig stolt over arbeidet mitt og håper den faller i smak hos deg også.

God lesing!

Bergen, mai 2023

Innholdsfortegnelse

Liste over tabeller og figurer

Innledning

Kunstig intelligens er et omfattende fagfelt som er i stadig utvikling, også innenfor utdanning. Innføringen av kunstig intelligente systemer i utdanningssektoren gjør det svært interessant å undersøke emnet. Før systemene kan implementeres fullstendig, er det nødvendig å forstå studentenes holdninger til å ta i bruk teknologien. I forlengelsen av teori og tidligere forskning som er redegjort i oppgavens teoretiske forankring, er formålet med denne studien å undersøke studenters tillit til å anvende kunstig intelligens i en vurderingsprosess. Som et resultat av dette er oppgavens problemstilling formulert som følgende:

I hvilken grad har studenter tillit til at en kunstig intelligens kan være rettferdig i en evalueringsprosess?

Det er viktig å påpeke at selv om denne teknologien ennå ikke er fullstendig utviklet, er formålet med studien å undersøke studenters holdninger og oppfatninger knyttet til potensialet for å bruke kunstig intelligens i en fullverdig evalueringsprosess i fremtiden. Ved å undersøke studenters respons på denne tematikken, oppnås innsikt i deres perspektiver. Det er viktig å understreke at denne studien ikke omfatter effektiviteten eller implementeringen av kunstig intelligens i dagens skolesystem, men heller å måle hvorvidt studenter har tillit til at systemet kan være rettferdig. Studien er også avgrenset til å ikke omfatte juridiske og teknologiske aspekter med systemet, samt ansvarsplassering ved eventuelle feil.

Forskning viser at kunstig intelligente systemer har større treffsikkerhet i vurderingen av flervalgsoppgaver og oppgaver med en tydelig fasit, sammenlignet med reflekterende oppgaver hvor fasiten er mer fleksibel (González-Calatayud et al., 2021). Med dette som utgangspunkt, er studien delt i to hovedkategorier basert på fagområder og studieretninger. Det skilles mellom fasitbaserte/konkrete fag som økonomi, matematikk og fysikk, og reflekterende/drøftende fag som psykologi, HR og filosofi. Det tas høyde for at fasiten i begge retningene betraktes som objektiv, og at mange faktorer spiller inn for å sette en karakter. Uavhengig av fag, er det viktig at alle eksamener vurderes på skjønn. På grunnlag av teorien om ulik treffsikkerhet hos kunstig intelligente systemer i de to fagområdene, ønsker denne studien å undersøke

om studenter har ulik grad av tillit til å bruke kunstig intelligens innenfor de ulike fagområdene. Dermed skiller studien mellom *fasitbaserte fag* og *reflekterende fag*.

1.0 Teoretisk forankring

1.1 Kunstig intelligens

Kunstig intelligens referer til teknologiske systemer som evner å utføre intelligente handlinger basert på store mengder data og en viss grad av autonomi, i den hensikt å oppnå et gitt mål (European Commission, 2019, s. 1-9). Systemene er utviklet av mennesker, for å behandle data og håndtere komplekse oppgaver. De kan ha menneskelignende egenskaper som: syn (bildegjenkjenning), hørsel (talegjenkjenning) og språk (naturlig språkgenerering). Kunstig intelligens anvendes blant annet i helsesektoren for å analysere sykdommer og stille diagnoser (Arnold, 2021, s. 121-139), for å forbedre offentlige tjenester (Asaro, 2019, s. 40-53), for å oppdage og forhindre forsikringssvindel (Aslam et al., 2022, s. 2-9) og for å styrke avlinger innenfor jordbruk (Young, 2020, s. 45-47). Videreutviklingen av data- og informasjonsbehandlingssystemer har ført til at kunstig intelligens også kan benyttes innenfor utdanningsfeltet for å veilede lærere, støtte elever og evaluere studenters arbeid (Ouyang, 2021, s. 1-5).

Utviklingen av kunstig intelligens går tilbake til 1950-tallet, da den engelske matematikeren Alan Turing formulerte Turingtesten. Testen skulle vurdere om en maskin kunne overbevise et menneske til å tro at den er menneskelig (Dvergsdal & Karlsen, 2023). På 1980-tallet ble konseptet om ekspertsystemer utviklet. Ekspertsystem er et datasystem hvor kunstig intelligens benyttes for å simulere beslutningsprosessen til en ekspert, innenfor et avgrenset fagområde (Tidemann, 2021a). Maskinlæring og dyplæring ble introdusert på 1980- og 1990- tallet. Maskinlæring handler om at datamaskiner får evner til å lære og forbedre seg selv basert på erfaringer og store mengder data, uten å bli eksplisitt programmert (Tidemann & Elster, 2022). Datamaskinen Deep Blue vant verdensmesterskapet i sjakk i 1997, og viste for alvor potensialet til kunstig intelligente roboter (IBM, u.å.a). Det ble forsket mye på nevrale nettverk på 1990- og 2000- tallet, men gjennombruddet skjedde ikke før i 2012 (Tidemann, 2023b). Etter dette har nevrale nettverk vært en dominerende metodikk innenfor kunstig intelligens. Den senere

tiden har utviklingen akselerert kraftig, og teknologien blir stadig mer avansert. I dag er kunstig intelligens et viktig fagfelt som har stor innvirkning på flere områder i hverdagen vår, fra å påvirke kjøpsmønstre til å forenkle lærer- og studiehverdagen.

Digitale assistenter som Siri (Apple) og Alexa (Amazon) er eksempler på vellykkede kunstig intelligente roboter. De evner å besvare spørsmål og utføre handlinger knyttet til systemets funksjoner. Systemene gir enkel tilgang på informasjon, er tilpasningsdyktige og kan bidra til å forbedre tilgjengeligheten, spesielt for brukere med nedsatt funksjonsevne. På denne måten kan kunstig intelligens ha en positiv effekt, men det finnes også eksempler hvor kunstig intelligens ikke fungerer like bra. I 2011 lanserte den amerikanske teknologibedriften IBM en av verdens mest omfattende kunstig intelligente plattformer, IBM Watson (IBM, u.å.b). IBM Watson Health utviklet tjenesten Watson for Oncology, som lenge ble omtalt som fremtiden innenfor helsetjenester. Den kunstig intelligente roboten var designet for å hjelpe leger og sykepleiere, ved å gi informasjon om effektiv behandling for kreftpasienter. Problemet var at roboten ble trent på teoretiske data fra et lite utvalg, og ikke ekte pasienter. Mangel på informasjon om virkelige pasienttilfeller og deres behov, førte til at Watson ga feil prediksjoner og anbefalinger (Konam, 2022). Dette viser viktigheten av å sørge for at systemene er grundig testet før de tas i bruk.

Big data er datasett som er så store og komplekse at de er utfordrende å behandle med tradisjonelle analyseteknikker. Big data og kunstig intelligens har en synergisk effekt: kunstig intelligens krever store mengder data for å forbedre beslutningsprosesser, og big data anvender kunstig intelligens for å analysere informasjon (Qlik, 2023). Kombinasjonen av kunstig intelligens og big data gir innovative verktøy og avanserte analysefunksjoner, som er nødvendig for å oppnå grundig innsikt i dataene. Innenfor utdanning gir big data muligheten til å samle inn og analysere store mengder data for å oppdage mønster og sammenhenger. Dette gir innsikt i studentenes læring, undervisningsmetoder og utdanningssystemene. Analyser av big data kan identifisere individuelle behov og utfordringer slik at læringsprosessen kan tilpasses hver enkelt student (Klašnja-Milićević et al., 2017, s. 1066-1078).

Den høye treffsikkerheten hos enkle algoritmer forklares av to faktorer: bias og noise (Kahneman, Rosenfield, Gandhi, & Blaser, 2016). Bias gjør det enkelt for en

statistisk algoritme å både inkludere og ekskludere relevante faktorer, som for eksempel kjønn, utseende og personlige preferanser. Dette står i kontrast til menneskehjernen, som lettere lar seg påvirke av all tilgjengelig informasjon (Sjåstad, 2019, s. 63-79). Et eksempel på dette er hvordan mennesker påvirkes av personlige preferanser til andre medarbeidere, og derfor kan inkluderer både relevant og irrelevant informasjon i en ansettelsesprosess. Algoritmenes evne til å utelukke irrelevante faktorer viser dermed hvordan algoritmene kan ha en høyere treffsikkerhet. Den andre faktoren som forklarer treffsikkerheten til algoritmer, er *noise* (støy). Dette handler om at en statistisk algoritme vil komme frem til det samme resultatet hver gang, gitt evalueringskriterier og all informasjon tilgjengelig (Sjåstad, 2019, s. 63-79). Altså, at det er mindre tilfeldig variasjon. En studie av Kahneman (2011), viser hvordan erfarne radiologer evaluerte det samme røntgenbilde to ganger, og i 20% av tilfellene gav de ulike resultater til det samme røntgenbildet. Dermed kan støyproblemet reduseres dersom menneskelige vurderinger erstattes med algoritmer som fatter en egen beslutning basert på dataene den er gitt.

Kunstig intelligens kan kategoriseres som enten sterk eller svak. Inndelingen kommer fra Searles` skille mellom *å være* intelligent, og *å oppføre seg* intelligent (Elements Of Ai, u.å.). Sterk kunstig intelligens ligner menneskelig intelligens og oppførsel. Denne teknologien har en selvbevissthet, men er ikke optimalt utviklet ennå. Det som er utviklet til nå, er svak kunstig intelligens. Svak kunstig intelligens kan ikke tenke selv, men evner å utføre oppgaver i sanntid gjennom et gitt datasett. På denne måten bidrar svak kunstig intelligens til å øke effektivitet og produktivitet.

1.2 Kunstig intelligens i utdanningssektoren

Den digitale utviklingen har akselerert kraftig de siste årene, også innen utdanning. Overgangen til digital hjemmeskole under coronapandemien i 2020 og introduksjonen av chatGPT i 2022, er eksempler på dette. Utviklingen av kunstig intelligens har åpnet for nye muligheter innenfor utdanningssektoren, og har potensialet til å medføre langsiktige effekter på studentenes læring og utvikling. Ved å utnytte teknologien kan utdanningsinstitusjoner tilpasse undervisningen for å imøtekomme individuelle behov, begrense «drop out», og effektivisere evalueringsprosessen (Schiff, 2022, s. 527-563). Innføringen av kunstig intelligens i

utdanningssektoren innebærer også utfordringer, som for eksempel manglende menneskelig interaksjon, personvern og risikoen for bias. Et kritisk perspektiv til kunstig intelligens i utdanning, er bekymringen for avhengighet av teknologien. Det er en risiko for at studenter kan bli avhengig av teknologien for å løse problemstillinger. På denne måten kan deres evne til å tenke kritisk begrenses. Videre reiser innføringen etiske spørsmål knyttet til ansvar og rettferdighet, og konsekvenser av å potensielt erstatte lærere. Implementeringen av kunstig intelligens krever gjennomtenkte strategier og retningslinjer som bevarer god opplæring blant både lærere og studenter, samt transparent og ansvarlig bruk av teknologien. Tiltak for dette blir redegjort sammen med bias, black box-problemet og algoritme aversjon senere i studiens teoridel.

Kunstig intelligens, nevrale nettverk, maskinlæring og big data har utviklet automatiserte vurderingssystemer til bruk i utdanning (Ramesh & Sanampudi, 2022, s. 2495-2527). For øyeblikket har disse systemene større treffsikkerhet i vurderingen av enkle oppgaver med en konkret fasit, sammenlignet med skriftlige, reflekterende oppgaver (González-Calatayud et al., 2021). Teknologien bidrar til at studenter kan få raskere tilbakemelding på vurderinger, og er tid- og ressursbesparende for lærere. Det er likevel flere utfordringer før denne teknologien kan innføres fullstendig. Kunstig intelligens og maskinlæring trenes på store mengder tidligere data, og det stilles spørsmål til hvilke tidligere oppgaver som skal brukes for å kalibrere og teste systemet. Videre blir det også reist spørsmål knyttet til selve poenggivningen, blant annet hvilke poengsummer som skal legges til grunn, og hvem som skal være ansvarlig for å fastsette disse (Raczynski & Cohen, 2018, s. 233-240). Disse spørsmålene blir ikke besvart i denne studien, men det viktig å merke seg at denne problematikken eksisterer.

Regjeringens strategi for digital kompetanse og infrastruktur i barnehage og skole inkluderer rammer, regelverk og strategier for å håndtere den digitale utviklingen i Norge. I strategien uttaler regjeringen at "digitalisering endrer oss, enten vi vil eller ikke. Men vi skal være med på å styre endringen, og vi må ha som mål å ligge i forkant." (Kunnskapsdepartementet, 2023, s. 3). Med dette ønsker regjeringen å legge til rette for å følge utviklingen, ved å sikre gode strukturer og kompetanse på

området. Digitaliseringen er sannsynligvis den mest omfattende endringen i norsk skole de siste årene, og regjeringen ønsker å ta i bruk digitale løsninger i større grad. Forutsatt at tjenestene bidrar til å øke kvaliteten på utdanningstilbudet og elevenes læring. Dette viser at regjeringen forventer, og støtter, at kunstig intelligente systemer blir en omfattende del av utdanningssektoren.

Roll og Wylie (2016) har identifisert flere aspekter som peker mot et paradigmeskifte i den tradisjonelle utdanningsmodellen. Kunstig intelligente systemer gjennomgår en betydelig utvikling, og vil trolig fortsette å bli implementert i økende grad. Til tross for at teknologien ennå ikke er fullstendig utviklet og i bruk, indikerer forskning at teknologien med stor sannsynlig vil ha en betydelig innvirkning på hverdagen vår. Dermed ønsker denne studien å undersøke studenters holdninger til muligheten for fremtidig implementering.

1.3 Tillit til kunstig intelligens

Tillit kan defineres som en opplevelse at andre mennesker eller institusjoner er trygge å samhandle med (Conradsen, 2022). Tillit til en kunstig intelligens handler om å ha tillit til at systemet har den nødvendige teknologien som kreves for å oppfylle de formålene det er gitt. Glikson and Woolley (2020) definerer tillit til kunstig intelligens, som tendensen til å ta en vesentlig risiko fordi man tror på en større sannsynlighet for et positivt utfall. Tillit innebærer en emosjonell og psykologisk komponent som i teorien ikke kan tilegnes noe ikke-menneskelig. Systemene er basert på programvarer og data, og kan dermed ikke tilskrives menneskelige egenskaper som tillit. Ryan (2020) hevder derfor at det er mer hensiktsmessig å snakke om pålitelighet, istedenfor tillit når det kommer til kunstig intelligens. Pålitelighet handler om konsistens, og forventningen om at systemet utfører handlinger og oppgaver på den mest hensynsfulle måten. Dermed er pålitelighet avgjørende, for å bygge tillit til systemet.

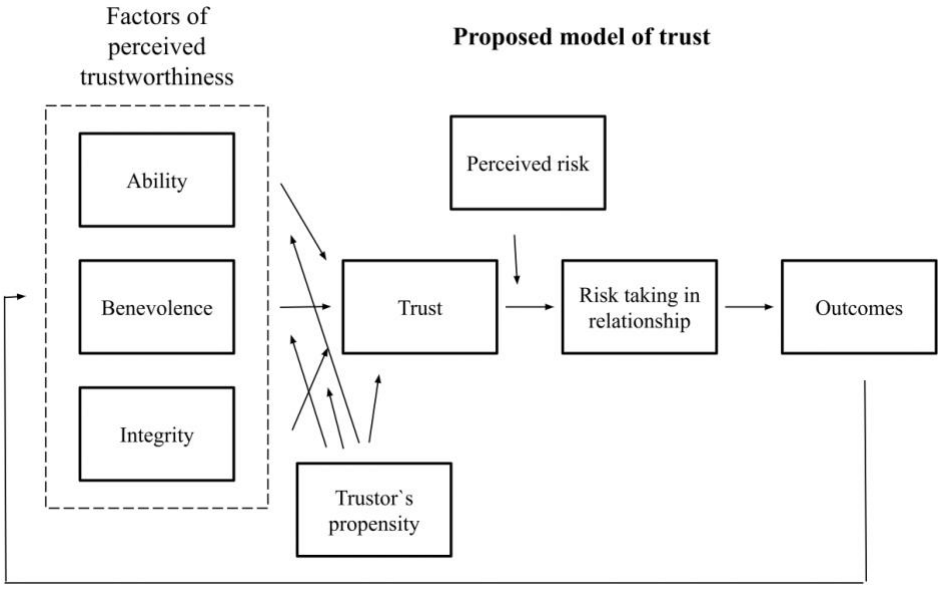
Validitet er en suksessfaktor for å oppnå tillit til en kunstig intelligens. Aloisi (2023) legger frem tre utfordringer ved validiteten til kunstig intelligente systemer i en evalueringsprosess: upålitelighet, lav forklaringsevne og bias. 1) Upålitelighet kan oppstå når to elever med tilsvarende besvarelser, får ulike karakterer. Kunstig intelligente evalueringssystemer kan ha, spesielt, stor grad av upålitelighet når det

kommer til å vurdere oppgaver innenfor reflekterende fag. Evalueringer av eksamener som etterspør åpne svar, krever i større grad fagkompetanse og en akademisk vurdering for å skille prestasjoner, noe en kunstig intelligens ikke evner. Mindre modifikasjoner som å endre et begrep, kan avgjøre hvordan en setning tolkes og på denne måten skape upålitelighet i systemet. 2) Lav forklaringsevne skyldes det såkalte black box-problemet, der algoritmene ikke kan gi en begrunnelse for resultatet som ble produsert. Selv om kunstig intelligente systemer evner å håndtere korte, enkle, besvarelser er ikke systemet nødvendigvis i stand til å forklare hvordan algoritmene kom frem til resultatet. Systemene mangler evnen til å forklare hvilke elementer som er vektlagt positivt eller negativt i evalueringen. 3) Bias referer til systematisk skjevhet som kan resultere i at studenter blir urettferdig vurdert. En svakhet ved automatiske evalueringssystemer er risikoen for at det blir satt feil karakter, som et resultat av bias.

Et annet aspekt ved tillit til kunstig intelligens er sikkerhet og personvern. Mennesker har ulik oppfatning av tillit, men tillit er avgjørende for å ta i bruk teknologien (Qin et al., 2020, s.1693-1710). Brukerne må kunne stole på at deres personlige data behandles på en sikker måte, og ikke misbrukes. Ved å formulere etiske retningslinjer for å beskytte personvern, etableres et grunnlag for ansvarlig og pålitelig bruk av systemene. Dette øker sannsynligheten for å oppnå tillit blant brukerne, og samfunnet som helhet. For å bygge tillit til kunstig intelligente systemer er det nødvendig å skape transparens i algoritmene, samt vise ansvarlighet. Transparens handler om å gi brukere innsikt i hvordan systemene fungerer, hvilke data som brukes og hvordan beslutninger tas. Transparens er med på å vurdere om systemet er rettferdig og pålitelig, og er derfor avgjørende for å skape tillit til kunstig intelligens. Ansvarlighet handler om at systemets utviklere skal stille seg bak teknologien og stå ansvarlig for eventuelle negative følger fra systemet. Disse komponentene er grunnleggende for å skape tillit til kunstig intelligens.

Mayer, Davis og Schoorman (1995) har konstruert en av de mest kjente modellene for å forstå tillit. Modellen består av tre grunnleggende faktor som påvirker tillit: kompetanse, integritet og velvilje. *Kompetanse* handler om at systemet har nødvendig kunnskap, ferdigheter og kapasitet til å oppnå ønskede resultater. I forbindelse med

tillit til kunstig intelligens innebærer dette at systemet har den nødvendige teknologien til å utføre handlinger, og ta beslutninger på en pålitelig måte. *Integritet* innebærer at systemet opptrer på en ærlig, pålitelig og rettferdig måte. Det handler om at systemet opererer i tråd med etiske retningslinjer og prinsipper. Kunstig intelligente systemer har integritet hvis det tar hensyn til transparens, rettferdighet og personvern. *Velvilje* handler om at systemets tar systemet tar beslutninger som er til fordel for brukeren. Tillit til kunstig intelligens innebærer tillit til at systemet er designet med en intensjon om å skape verdi og nytte i brukerens favør. Kompetanse, integritet og velvilje er gjensidig avhengig av hverandre og gir en synergisk effekt når det gjelder å påvirke tilliten til kunstig intelligens.



Figur 1: Proposed model of trust

1.4 Bias

Bias referer til systematiske skjevheter som påvirker resultatet i en uønsket retning (Grønmo, 2020). Bias omfatter urettferdig eller diskriminerende behandling basert på faktorer som kjønn, etnisitet og personlige preferanser. I kunstig intelligens oppstår bias som en konsekvens av at algoritmer og maskiner er trent på historiske data som inneholder skjevheter. Som et resultat av dette kan systemet ta ikke-objektive beslutninger, som strider med etiske prinsipper (Hall & Ellis, 2023). Skadelig bias har allerede blitt observert i kunstig intelligente systemer ved flere tilfeller. Eksempelvis,

ble det oppdaget at Googles annonseverktøy for målrettet annonsering, viste en betraktelig høyere andel annonser for høyt lønnede jobber til menn, sammenlignet med kvinner (Datta et al., 2015, s. 92-112). Dette avslørte en kjønns-diskriminerende form for bias, ved at annonseplattformen distribuerte og eksponerte jobbannonser basert på kjønn. Et annet eksempel hvor et kunstig intelligent system inneholdt skadelig bias, er Amazon sitt rekrutteringsverktøy som ble utviklet i 2018 (Dastin, 2018). Systemet var trent på tidligere CV-er og jobbsøknader, med den hensikt å vurdere nye søknader. Det viste seg at systemet favoriserte mannlige søkere, og på denne måten diskriminerte kvinner. Årsaken til skjevheten, var at de tidligere treningsdataene inkluderte en større andel mannlige søkere slik at systemet lærte seg å foretrekke menn.

Bias kan altså være svært alvorlig, også i en evalueringsprosess. Algoritmer som er preget av bias kan favorisere eller diskriminere visse grupper av studenter, til tross for at de har samme faglige kompetanse. Urettferdig behandling som følge av bias kan skape mistrivsel, frustrasjon og mistillit. En studie av Baker og Hawn (2022) identifiserer etnisitet, kjønn og nasjonalitet som de mest hyppige formene for bias i utdanningssektoren. Den samme studien inkluderer strategier og anbefalinger for å redusere bias i utdanning: 1) Representative treningsdata som inkluderer informasjon om sosioøkonomiske faktorer, demografi og kultur, 2) Reguleringer og etiske retningslinjer som definerer hvordan systemet brukes, 3) Transparens i algoritmene skaper forståelse for hvordan algoritmene fungerer, og 4) Kontinuerlig menneskelig vurdering av resultatet til algoritmene er med på å identifisere eventuelle avvik og korrigere skjevhet, samt vise ansvarlighet overfor systemet. Selv om det ikke er sikkert at bias noen gang kan elimineres fullstendig, har hvert av disse tiltakene likevel potensialet til å redusere omfanget av bias. Tiltak for å redusere bias kan bidra til å styrke tilliten til det kunstig intelligente systemet.

1.5 Black box-problemet

Black box-problemet referer til situasjoner hvor det er utfordrende å forstå hvordan kunstig intelligente systemer har kommet frem til resultatet. Selv om det er kjent hvilke data og informasjon som er gitt, og hva som er resultatet, forblir selve beslutningsprosessen ukjent (Gràcia & Sancho-Gil, 2021). Dataene i algoritmene er

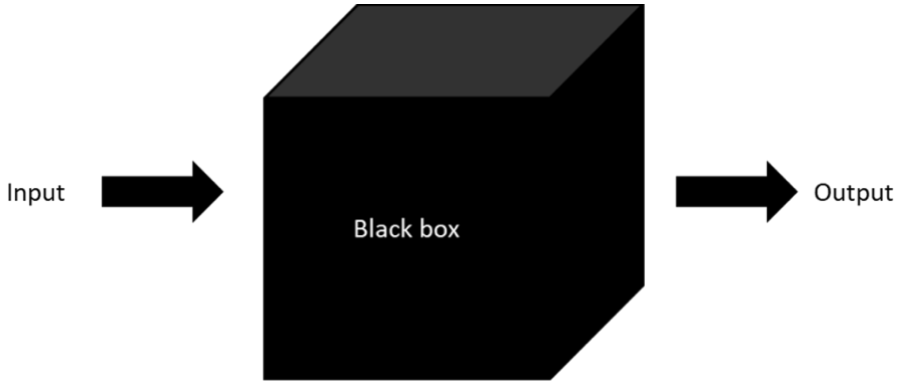
basert på tilgjengelige data og er derfor aldri objektive eller komplette. Som et resultat av dette kan datasettene unngå å inkludere betydningsfull informasjon som er nødvendig for å gi en beslutning som et menneske ville akseptert som det korrekte valget. Black box-modellen gjør det også utfordrende å identifisere bias som påvirker beslutningene, og gjør det dermed utfordrende å avdekke og håndtere slike skjevheter (Gillani et al., 2023, s. 99-111). I en britisk studie ble ansiktsgjenkjennings-teknologi brukt for å identifisere ansiktene til mørkhudede kvinner. Studien identifiserte kritiske diskriminerende bias knyttet til systemets eksisterende data (Buolamwini & Gebru, 2018). Manglede informasjon om hvordan systemet tar beslutninger kan medføre usikkerhet og bekymringer knyttet til ansvar og rettferdighet i systemet. Dette kan være spesielt problematisk i domener hvor feil beslutninger har betydningsfulle konsekvenser, for eksempel i helsevesenet og rettsvesen.

Til tross for algoritmenes høye treffsikkerhet, er det problematisk at beslutningsprosessen er ukjent. Dette gjør at mennesker ikke kan forstå hvorfor maskinen kommer fram til akkurat denne beslutningen. Et eksempel på dette er bruken av algoritmer i en låneprosess. Her er det ønskelig å få en begrunnelse på hvorfor noen får lån og noen ikke får det. Selv om mange hevder at maskinene er mer nøyaktig enn mennesker, vil det fremdeles oppstå feil som det er problematisk å ikke kunne forklare. Kunstig intelligens vil heller aldri kunne ta en beslutning som er 100% sikker, og det er derfor viktig å vurdere graden av usikkerhet (Øye & Normann, 2021). Feilaktige beslutninger kan være svært problematisk, og kan potensielt medføre alvorlige ulykker. For eksempel, kan dette være spesielt utfordrende i sammenheng med selvkjørende biler.

En studie gjennomført av Brown et al. (2017) demonstrerte hvordan tillegg av en ekstra detalj i et bilde av en banan, kunne forvirre en nevral-nettverksmodell som ble brukt til bildegjenkjenning. I studien ble det utført en eksperimentell manipulasjon der ett klistermerke av en brødrister ble lagt til ett bilde av en banan, noe som medførte at maskinen feilaktig klassifiserte bildet som en brødrister. Studien viser problematikken rundt black box-modellen, ved å demonstrere hvor enkelt algoritmer kan forvirres til å gi et feilaktig resultat. Overført til utdanningssektoren kan dette medføre betydelige konsekvenser. Eksempelvis kan dette gjøre at et automatisert

vurderingssystem favoriserer visse begreper og uttrykk i en besvarelse, og "straffer" studenter som ikke har inkludert dette. Manglende innsikt i hvilke faktorer som vurderes for å komme frem til et resultat kan skape usikkerhet blant brukerne. Tillit og rettferdighet har stor betydning i utdanningssektoren og black box-problemet understreker behovet for transparens, for å forstå hvordan algoritmene fungerer i en vurderingsprosess.

I en studie av Lipton (2017) ble det identifisert flere tiltak som kan bidra til å løse black box-problemet: 1) *Økt transparens* i algoritmene er med på å øke forklarbarheten til modellen, og gjør det også mulig å identifisere og korrigere eventuelle bias som oppstår, 2) *Økt innsats innen forskning* bidrar til å skaffe mer informasjon for å bedre forstå problemet og identifisere løsninger, og 3) *Etiske retningslinjer og reguleringer* som fremmer ansvarlighet, og sikrer at systemene er utviklet og brukes på en rettferdig måte. Det er også viktig å skape større bevissthet rundt black box-problemet, slik at flere arbeider for å redusere problematikken. I tillegg er det nødvendig å inkludere et brede mangfold i treningsdataene, slik at bias kan reduseres.



Figur 2: Black box-modell

1.6 Algoritme aversjon

Til tross for at forskning tyder på at algoritmer konsekvent er mer treffsikre enn menneskelige beslutninger er mange skeptiske til å ta dem i bruk, et fenomen kjent som algoritme aversjon (Dietvorst et al., 2018, s. 1155-1170). Algoritme aversjon oppstår når man har større tillit til at mennesker kan ta en rettferdig beslutning, fremfor beslutninger tatt av maskiner eller algoritmer (Prahl & Swol, 2017, s.691-

702). En studie gjennomført av Dietvorst et al. (2015) avslører at personene som observerte algoritmens utførelse, hadde en lavere tillit til den og viste mindre sannsynlighet for å velge den fremfor et mindre dyktig menneske. Dette var tilfellet også blant de som observerte at algoritmen presterte bedre enn mennesket. En årsak til dette kan være manglende tillit til teknologien, frykten for bias eller manglende transparens i algoritmenes funksjon. Dermed kan algoritme aversjon forklares som en skepsis til å bruke algoritmer, i frykt for å oppleve mangel på kontroll og potensiell urettferdighet.

Einhorn (1989) hevder at usikkerheten til algoritmer ikke nødvendigvis handler om at algoritmen vil gjøre feil, men at mennesker evner å utføre med perfeksjon. I tillegg viser en studie av Dietvorst et al. (2015), at mange er mindre tolerante for algoritmens *mindre* feil, enn menneskets *større* feil. Gitt at motviljen for ufullkomne algoritmer kommer av frykten for ukontrollerbare feil, kan man være mer åpne dersom det er mulig å redusere eller eliminere slike feil (Dietvorst et al., 2018, s. 1155-1169). Dette viser større villighet til å bruke ufullkomne algoritmer dersom det er mulig å kontrollere algoritmenes prognoser. På denne måten kan algoritme aversjon reduseres, til tross for at slike inngrep og justeringer ofte gjør algoritmene dårligere.

En studie av Reich et al. (2022) avdekket behovet for å implementere tiltak som kan bidra til å redusere skepsisen til algoritmene. Studien hevder at økt transparens i algoritmene er en sentral driver for å redusere algoritme aversjon. Ved å gjøre algoritmene mer forklarbare, kan man bygge tillit ved å gi brukerne innsikt i prosessen. Økt kunnskap og bevissthet blant brukerne, kan også bidra til å redusere skepsisen for å bruke algoritmene. Videre understreker studien betydningen av etiske retningslinjer er nødvendig for å redusere bias og opprettholde en høy kvalitet i systemet.

1.7 Hypoteser

For å besvare problemstillingen er det utarbeidet 9 hypoteser som skal undersøkes og analyseres. Hypotesene er formulert på grunnlag av tidligere forskning og relevant teori som er redegjort i oppgavens teoretiske del. På denne måten sikres en teoretisk forankring, og en sammenheng i studien. Sentrale begreper som er relevante for

problemstillingen danner grunnlaget for variablene som skal analyseres gjennom hypotesene. Følgende hypoteser er utviklet:

H1: Alder har en negativ effekt på tillit

H2: Fasitbaserte fag har en positiv effekt på tillit

H3: Bias har en negativ effekt på tillit

H4: Black box-problemet har en negativ effekt på tillit

H5: En hypotetisk eliminering av black box-problemet har en positiv effekt på tillit

H6: En hypotetisk eliminering av algoritme aversjon har en positiv effekt på tillit.

H7: Gjennomsnittlig tillit er høyere i fasitbaserte fag sammenlignet med reflekterende fag

H8: Studenter har i gjennomsnitt tillit til at en kunstig intelligens kan være rettferdig i en evalueringsprosess

H9: Det er en sammenheng mellom holdningen til eliminering av algoritme aversjon og tillit

2.0 Metode

Metode refererer til den systematiske fremgangsmåten som brukes for å samle inn, analysere og tolke data og informasjon (Johannessen et al., 2011, s. 29). I denne delen av oppgaven blir det gjort rede for undersøkelsens forskningsdesign, utvalg og fremgangsmåten for eksperimentet, samt gjennomføring av datainnsamling og valg av metode.

2.1 Forskningsdesign

Forskningsdesign er en overordnet plan for hvordan en undersøkelse skal gjennomføres. Det inkluderer blant annet hvem og hva som skal undersøkes, og hvordan problemstillingen skal belyses og besvares (Johannessen et al., 2011, s. 73). Formålet med studien er å løse problemstillingen *i hvilken grad har studenter tillit til at en kunstig intelligens kan være rettferdig i en evalueringsprosess*. For å finne svar på dette innhentes data ved bruk av både kvantitativ og kvalitativ forskningsmetode.

En kombinasjon av kvantitativ og kvalitativ metode gir et helhetlig bilde av problemstillingen. Kvantitativ metode gir et større perspektiv, mens kvalitativ metode går mer i dybden. Variabelen som undersøkes er *tillit* til kunstig intelligens, og måles ved hjelp av data innhentet fra spørreundersøkelse og intervjuer.

2.1.1 Kvantitativ metode

Kvantitativ metode er en forskningsmetode som omhandler innsamling og analyse av målbare data. Kvantitative data uttrykkes i form av tall, og analyseres ved bruk av statistiske analyser (Grønmo, 2023b). Metoden involverer ofte bruk av strukturerte spørreundersøkelser eller meningsmålinger, for å samle inn data fra en større mengde respondenter på kort tid. Formålet er å samle inn data fra et utvalg som kan generaliseres til en større populasjon, slik at man kan identifisere sammenhenger og trekke konklusjoner på et større nivå. Oppgavens kvantitative metode er en digital spørreundersøkelse. En svakhet ved metoden er at spørreundersøkelsen er begrenset til forhåndsdefinerte svaralternativer, noe som kan hindre deltakernes mulighet til å uttrykke sine faktiske tanker om fenomenet. For å kompensere for dette, benyttes kvalitativ metode som et supplement for kvantitativ metode.

Spørreundersøkelsen defineres som en kvantitativ undersøkelse med både et deskriptivt design og et tverrsnittsdesign. Et deskriptivt design har som formål å beskrive og kvantifisere egenskaper ved en populasjon eller utvalg (Johannessen et al., 2011, s. 415). Med andre ord, et deskriptivt design har som mål å beskrive "hvordan ting er" og ikke "hvorfor ting er som de er". Undersøkelsen kategoriseres også som en tverrsnittsundersøkelse, ettersom den samler inn data fra ett, og kun ett, tidspunkt (Johannessen et al., 2011, s. 74). Individers tillit til kunstig intelligens kan endres over tid, og undersøkelsen gir derfor et øyeblikksbilde av fenomenet.

Spørreundersøkelsen er utformet og utviklet gjennom det digitale programmet Qualtrics, og deltakerne gjennomførte undersøkelsen i samme program. Eksperimentets målgruppe er studenter, og undersøkelsen ble distribuert til respondenter over privat melding. For å sikre et representativt utvalg ble undersøkelsen distribuert til et tilfeldig utvalg av både kvinner og menn i ulike aldersgrupper, uavhengig av studieretning og nivå på studiet. Ved starten av undersøkelsen ble alle deltakerne informert om anonymisering, og måtte gi sitt

samtykke for å kunne delta. Spørsmålene i spørreundersøkelsen er utformet med utgangspunkt i oppgavens teoretiske del, samt de hypotesene som skal analyseres. Dette bidrar til å opprettholde en helhetlig og gjennomgående oppgave.

2.1.2 Kvalitativ metode

Kvalitativ metode er en forskningsmetode som fokuserer på å analysere og tolke komplekse fenomener. Kvalitative data uttrykkes som oftest i form av tekst, og kan innhentes gjennom eksempelvis intervjuer og observasjoner. Formålet er å skaffe innsikt i hvordan mennesker opplever og forstår ulike situasjoner og fenomener (Grønmo, 2023a). I denne oppgaven er den kvalitative metoden basert på gjennomføringen av fire dybdeintervjuer med ulike studenter. Hensikten med disse intervjuene er å oppnå en dypere forståelse for forskningen, ved å gi detaljert innsikt i studentens synsvinkler og holdninger.

Intervjuene karakteriseres som kvalitative dybdeintervjuer med et semistrukturert design. Semistrukturerte intervjuer baserer seg på en overordnet intervjuguide med et forhåndsbestemt tema og forhåndsdefinerte spørsmål, samtidig som intervjueren står fritt til å stille oppfølgingsspørsmål og intervjuobjektet får mulighet til å utdype svarene sine (Johannessen et al., 2011, s. 137-139). Hensikten med denne intervjuformen er å innhente detaljert informasjon om intervjuobjektets synspunkter, holdninger og erfaringer om et bestemt tema. Semistrukturerte intervjuer identifiserer sentrale synspunkter knyttet til problemstillingen, og gir en dypere innsikt i forskningsområdet. Et semistrukturert intervjudesign gir stor grad av fleksibilitet og egner seg til komplekse problemstillinger.

Spørsmålene i intervjuene er utformet i samsvar med oppgavens teoretiske forankring, hypoteser og spørreundersøkelsen. På denne måten sikres koherens i forskningen. Ettersom intervjuene har et semistrukturert design er spørsmålene formulert på forhånd, og medbragt til intervjuene. Flertallet av spørsmålene er åpne, slik at intervjuobjektene har mulighet til å svare ærlig og utdypende om sine synspunkter. Intervjuene ble avholdt ansikt til ansikt, og intervjueren tok notater underveis. For å sikre validitet, holdt intervjueren en nøytral tilnærming til problemstillingen slik at intervjuobjektene ikke skulle bli påvirket. Totalt ble det gjennomført fire dybdeintervjuer for å sikre dybde i dataen i form av flere

synsvinkler. Ettersom studiens målgruppe er studenter, var samtlige intervjuobjekt studenter. For å ivareta personvernet til intervjuobjektene, overholdes personvernregler som GDPR. Dette innebærer anonymisering av intervjuobjektene i oppgavens skriftlige materiale, i samsvar med retningslinjene fra Datatilsynet (2019).

2.1.3 Primær- og sekundærdata

Primærdata er ny data som forskere selv samler inn for prosjektets formål (Ringdal, 2018, s. 124). Denne oppgavens primærdata omhandler rådata som er samlet inn gjennom den digitale spørreundersøkelsen, og dybdeintervjuene. En fordel med primærdata er at det kan være svært detaljert og nøyaktig, ettersom det er spesifikt innhentet til det aktuelle prosjektet. Dette gjør primærdata til verdifull informasjon som kan bidra til å løse problemstillingen og underbygge hypotesene i forskningen.

Sekundærdata referer til data som allerede er samlet inn av noen andre, til et annet formål (Ringdal, 2018, s. 118). Denne oppgavens sekundærdata omfatter tidligere forskningsdata, fagartikler og relevant litteratur. Dette bidrar til å validere og styrke forskningsresultatene ved å tilføre et bredere perspektiv i lys av problemstillingen. På denne måten er sekundærdata et verdifullt supplement til primærdataene, ved å bidra til å danne et mer komplett bilde av fenomenet.

2.2 Validitet og reliabilitet

Validitet handler om hvor troverdig dataene er, og i hvilken grad de måler det de faktisk skal måle (Johannessen et al., 2011, s. 230). I både spørreundersøkelsen og intervjuene referer validitet til at spørsmålene som stilles faktisk representerer det fenomenet som undersøkes. For å sikre validitet er spørsmålene formulert på bakgrunn av teorier og tidligere forskning som er framlagt i oppgavens teoretiske del. Dette gir en solid teoretisk forankring og bidrar til å opprettholde en kontinuerlig sammenheng gjennom hele studien, noe som styrker validiteten og reliabiliteten til resultatene. I tillegg bidrar intervjuerens nøytrale holdning til å sikre validitet. Ved å være nøytral og ikke påvirke intervjuobjektets svar, bidrar intervjueren til å opprettholde validiteten i datainnsamlingen.

Reliabilitet referer til graden av konsistens og pålitelighet i målingene som gjøres. Det handler om i hvilken grad man kan stole på dataene, og hvor nøyaktig de

representerer den aktuelle begivenheten som undersøkes (Johannessen et al., 2011, s. 40). For å sikre både reliabilitet og validitet i spørreundersøkelsen er spørsmålene nøye formulert på en presis og forståelig måte, og uten rom for tolkning. Spørsmålene ble testet på en pilotgruppe for å identifisere eventuelle misforståelser eller uklarheter. Videre er det sikret en tilstrekkelig utvalgsstørrelse, samt et representativt utvalg. Dette bidrar til å redusere sannsynligheten for tilfeldig variasjon. For å øke reliabiliteten i eksperimentet, blir det også utført en intern konsistensanalyse ved hjelp av Cronbachs alfa.

For å sikre reliabilitet i dybdeintervjuene benyttes et semistrukturert design, med en forhåndsbestemt intervjuguide og forhåndsbestemte spørsmål. På denne måten fungerer intervjuguiden som en veiledning under intervjuene, og sørger for at alle relevante områder blir dekket. Ved å anvende forhåndsdefinerte spørsmål sikres det at alle intervjuobjektene blir stilt de samme spørsmålene, noe som bidrar til konsistens og pålitelighet i datainnsamlingen. I tillegg er spørsmålene formulert på en klar og presis måte, og de har blitt testet gjennom en pilotgruppe for å identifisere eventuelle misforståelser eller uklarheter.

3.0 Resultater og funn

I denne delen av oppgaven presenteres resultater og funn som er relevant for problemstillingen. Dataene ble innhentet ved bruk av både kvantitativ og kvalitativ forskningsmetode, i form av en spørreundersøkelse og dybdeintervjuer med studenter. Totalt 302 respondenter startet på spørreundersøkelsen, der 235 fullførte hele. Av disse var 40,14% kvinner, og 59,86% menn. 41,96% tilhørte aldersgruppen 21-23, og aldersgruppe 18-20, 24-26, 27-29 og 30+ utgjorde henholdsvis 17,48%, 24,48%, 10,49% og 5,99% av utvalget. Resultatene ble behandlet i dataprogrammet JMP, hvor ulike statistiske analyser som regresjonsanalyse, t-tester, krysstabulering og kji-kvadrattest ble utført. Målet med analysene var å undersøke om hypotesene støttes eller forkastes. Gjennom dybdeintervjuene ble det innhentet utdypende informasjon fra fire studenter. Dybdeintervjuene har som mål å gi en dypere forståelse av deltakernes holdninger og synspunkter knyttet til problemstillingen, og funnene presenteres i en oppsummerende tabell for å gi en sammenfattende oversikt.

3.1 Cronbachs alfa

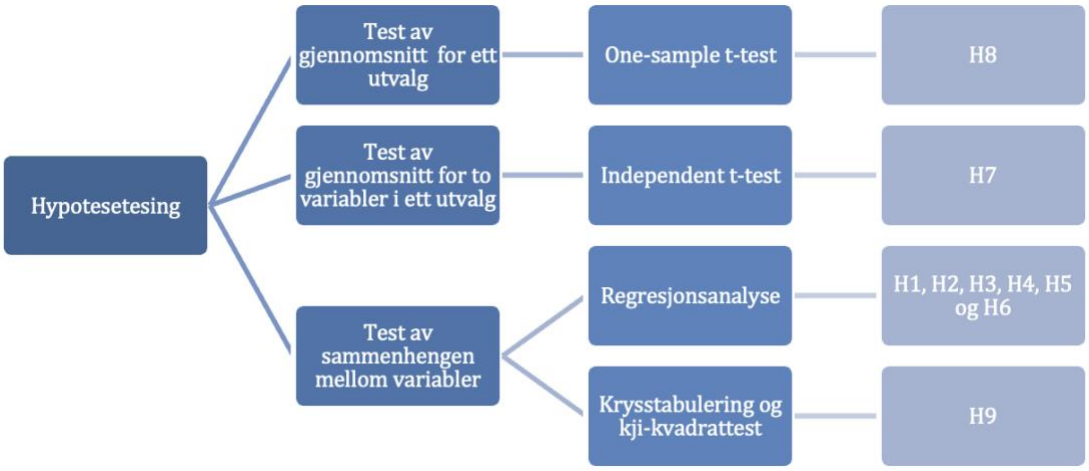
Cronbachs alfa er en statistisk metode for å måle reliabiliteten mellom spørsmålene i undersøkelsen (Silkoset et al., 2022, s. 257). Ved å vurdere den gjennomsnittlige korrelasjonen mellom spørsmålene beregnes en numerisk verdi mellom 0 og 1. En høy verdi av Cronbachs alfa indikerer at målingene i spørreundersøkelsen er pålitelige og konsistente. For å oppnå reliabilitet anbefales det at alfa verdien er over 0,7 (Ringdal, 2018, s. 104). Resultatene fra analysen viser en Cronbachs alfa på 0,71 for variablene som ble målt. Dette tyder på at målingene er relativt pålitelige og konsistente.

Variabler	**Cronbachs alfa**
Q5, Q6, Q7, Q8, Q9, Q10, Q11	0,7134

Tabell 1: Cronbachs alfa

3.2 Hypotesetesting

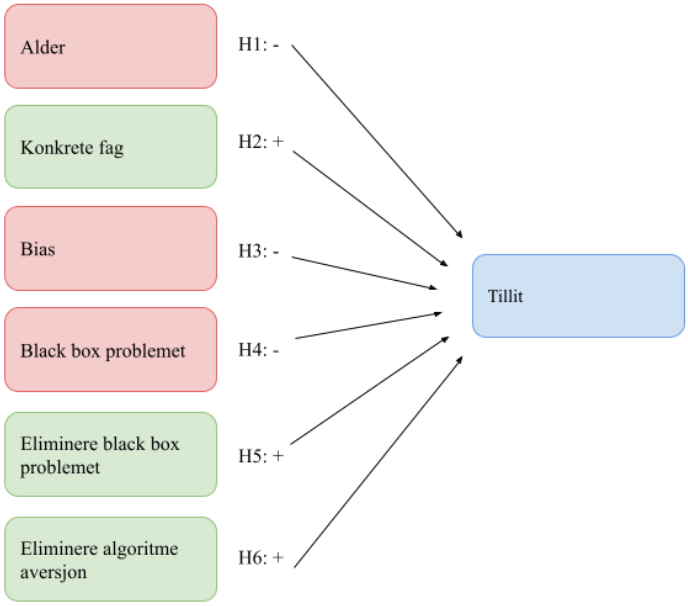
For å teste hypotesene anvendes flere ulike analytiske metoder, inkludert regresjonsanalyse, t-tester, krysstabulering og kji-kvadrattest. Ved å bruke flere ulike analytiske metoder oppnås en omfattende og helhetlig analyse av variablene som undersøkes. Dette bidrar til å belyse ulike aspekter og sammenhenger mellom variablene, og resulterer i en dypere forståelse av problemstillingen. Bruken av flere analytiske metoder styrker validiteten og reliabiliteten i forskningen, da sammenfallende funn på tvers av ulike analyser forsterker den overordnede konklusjonen. Videre kan de ulike analysene utfylle hverandre ved å avdekke forskjellige perspektiver og nyanser i dataene, noe som bidrar til å styrke den overordnede kvaliteten på studien. Ved å inkludere ulike analyser i forskningen presenteres sammenhengene mellom variablene på en mer pålitelig og nyansert måte.



Figur 3: Oversikt over de ulike hypotesetestene

3.2.1 Regresjonsanalyse

Regresjonsanalyse er en statistisk metode som brukes for å undersøke og evaluere sammenhengen mellom en avhengig variabel, *Y*, og flere uavhengige variabler, *X* (Silkoset et al., 2022, s. 356). En multippel regresjonsanalyse gir en omfattende forståelse av sammenhengen mellom variablene, samtidig som den tar hensyn til potensielle korrelasjoner. I tillegg identifiseres hvilke variabler som er mest relevant for å forklare variasjonen i den avhengige variabelen. Analysen ønsker å undersøke om det er en statistisk signifikant sammenheng mellom de uavhengige variablene *alder*, *fasitbaserte fag*, *bias*, *black box-problemet*, *eliminering av black box-problemet* og *eliminering av algoritme aversjon*, og den avhengige variabelen *tillit*. I samsvar med teori og forskning som er redegjort i oppgavens teoretiske del, samt egne antagelser, er det formulert ulike hypoteser om de forventede sammenhengene mellom variablene. Hypotesene er som følgende: H1: Alder har en negativ effekt på tillit, H2: Fasitbaserte fag har en positiv effekt på tillit, H3: Bias har en negativ effekt på tillit, H4: Black box-problemet har en negativ effekt på tillit, H5: En hypotetisk eliminering av black box-problemet har en positiv effekt på tillit og H6: En hypotetisk eliminering av algoritme aversjon har en positiv effekt på tillit. Hypotesene er oppsummert i modellen under.



Figur 4: Forskningsmodell over sammenhenger til tillit

Resultatene fra regresjonsanalysen er fremvist i tabellen under:

Avhengig variabel: Tillit					
Uavhengige variabler	**Ustandardisert regresjons-parameter**	**95% konfidensintervall**		**Standardisert regresjons-parameter**	**p-verdi**
		Nedre grense	**Øvre grense**		
Intercept	5,2	2,53	7,8	-	**
Alder	-0,14	-0,34	0,06	-0,08	NS
Fasitbaserte fag	0,55	0,15	0,94	0,18	**
Bias	-0,37	-0,7	-0,04	-0,16	*
Black box-problemet	-0,8	-1,16	-0,45	-0,31	***
Eliminering av black box-problemet	0,85	0,12	1,6	0,15	*
Eliminering av algoritme aversjon	0,02	-0,3	0,35	0,01	NS

Tabell 2: Resultater fra regresjonsanalysen

Note: R^2 = 15,4%, justert R^2 = 13%. NS = ikke signifikant. * $p < 0,05$; ** $p < 0,01$, *** $p < 0,001$

Regresjonsanalysen kommer frem til følgende regresjonslikning:

Tillit = 5,2 – 0,14alder + 0,55fasitbaserte fag – 0,37bias – 0,8black box + 0,85xblack box + 0,02algoritme aversjon.

Standardisert regresjonsparameter:

Standardisert regresjonsparameter gjør det mulig å sammenligne flere uavhengige variabler, uavhengig av deres målemetode (Silkoset, 2021, s. 373). Verdien indikerer både retningen og styrken ved hver variabel, i forhold til den avhengige variabelen. En positiv standardisert regresjonskoeffisient viser en positiv sammenheng, derimot viser en negativ koeffisient det motsatte. Resultatene fra regresjonsanalysen avslører at *black box-problemet* har den sterkeste påvirkningen, med en negativ standardisert regresjonsparameter på -0,31. *Bias* indikerer også en negativ verdi på -0,16. Imidlertid har *fasitbaserte fag* og *eliminering av black box-problemet* positive, standardiserte regresjonsparametere, på henholdsvis 0,18 og 0,15. Variablene *alder* og *eliminering av algoritme aversjon* er ikke signifikante og vurderes derfor ikke.

Rapportering fra regresjonsanalysen:

Hypotese 1 hevdet at alder har en negativ effekt på tillit. Regresjonsanalysen viser en negativ standardisert regresjonsparameter på -0,08. Hypotesen fikk ikke støtte fra p-verdien på 5% signifikansnivå, og blir dermed forkastet.

Hypotese 2 påsto at fasitbaserte fag har en positiv effekt på tillit. Resultatene fra regresjonsanalysen viser en positiv standardisert regresjonsparameter på 0,18, og en p-verdi på < 0,01. Dette gir støtte til hypotesen, og indikerer en signifikant positiv sammenheng mellom den uavhengige variabelen fasitbaserte fag og den avhengige variabelen tillit.

Hypotese 3 hevdet at bias har en negativ effekt på tillit. Analysen viser en standardisert regresjonsparameter på -0,16, som er signifikant forskjellig fra null på 5% signifikansnivå ($p < 0,05$). Dette støtter hypotesen, og det kan konkluderes med at bias har en negativ effekt på tillit.

Hypotese 4 påsto at black box-problemet har en negativ effekt på tillit. Resultatene fra analysen viser at black box-problemet har den sterkeste påvirkning av alle

variablene, med en standardisert regresjonsparameter på -0,31. Den ekstremt lave p-verdien på < 0,001 indikerer 99,9% sikkerhet for at sammenhengen mellom variablene er signifikant. Dette gir sterk støtte til hypotesen. Basert på analysen kan man si at det er betydelig bevis for at black box-problemet faktisk har en negativ effekt på tillit.

Hypotese 5 hevdet at en hypotetisk eliminering av black box-problemet ville hatt en positiv effekt på tillit. Resultatene fra regresjonsanalysen viser en standardisert regresjonsparameter på 0,15, og en p-verdi < 0,05. Dermed støttes hypotesen, og det konkluderes med at dersom black box-problemet elimineres vil det ha en positiv effekt på tillit. Det understrekes at denne hypotesen representerer et teoretisk scenario, noe som vil bli grundigere diskutert i drøftedelen av oppgaven.

Hypotese 6 påsto at en hypotetisk eliminering av algoritme aversjon ville hatt en positiv effekt på tillit. Resultatene fra analysen viser en standardisert regresjonsparameter på 0,01, men p-verdien gir imidlertid ikke støtte til hypotesen. Dermed forkastes hypotese 6, og det er ikke tilstrekkelig bevis for at det er en sammenheng mellom eliminering av algoritme aversjon og tillit.

Regresjonsmodellen har en forklaringskraft (R^2) på 15,4%, som betyr at 15,4% av variasjonen i den avhengige variabelen kan forklares av de uavhengige variablene. Dette indikerer at 84,6% av variasjonen ikke kan forklares av de inkluderte uavhengige variablene, og det kan være andre faktorer som påvirker tillit som ikke er tatt med i modellen. Den justerte R^2 –verdien på 13%, tar hensyn til modellens kompleksitet og tar hensyn til antall forklaringsvariabler.

Dersom de "ikke-signifikante" variablene utelukkes fra modellen, fås følgende ligning:

Tillit = 5,2 + 0,55fasitbaserte fag – 0,37bias – 0,8black box + 0,85xblack box

3.2.2 T-test

T-test er en statistisk metode som brukes for å undersøke om det er signifikant forskjell mellom gjennomsnittet i to variabler, eller mellom gjennomsnittet for en variabel og en hypotetisk verdi. Metoden tar hensyn til variasjonen i dataene og antall

observasjoner for å vurdere om den observerte forskjellen er statistisk signifikant, eller om det forskjellen kan tilskrives tilfeldigheter (Silkoset et al., 2022, s. 306-318).

En independent t-test brukes for å sammenligne gjennomsnittene mellom to uavhengige variabler i samme utvalg (Ringdal, 2018, s. 384-385). Testen brukes for å undersøke om det er en signifikant forskjell mellom gjennomsnittene i fasitbaserte og reflekterende fag. Følgende hypotese testes:

H7: Gjennomsnittlig tillit er høyere i fasitbaserte fag sammenlignet med reflekterende fag

t-verdi	**p-verdi**	**Cohens d**
18,49	< ,0001	1,16

Tabell 3: Resultater fra independent t-test

Note: antall frihetsgrader (DF) = 255

Resultatene fra t-testen viser en signifikant sammenheng [t (255) =18,49, p<0,0001, d=1,16] mellom variablene. Den høye t-verdien indikerer en betydelig forskjell i gjennomsnittene mellom variablene, og den ekstremt lave p-verdien indikerer at forskjellen ikke er tilfeldig. Cohens d er en effektstørrelse basert på standardavviket i variablene, og den måler størrelsen på forskjellen mellom gjennomsnittene. En d-verdi på 0,2 anses vanligvis som en liten effekt, 0,5 som en moderat effekt og 0,8 som en stor effekt (Cohen, 1988, s. 26). Cohens d = 1,16 indikerer dermed en betydelig forskjell mellom gjennomsnittene. Resultatene gir støtte til hypotesen, og det kan konkluderes med at studenter har større tillit til at kunstig intelligente systemer kan gi en rettferdig karakter i fasitbaserte fag, sammenlignet med reflekterende fag.

En one-sample t-test brukes til å vurdere om gjennomsnittet for en enkel variabel er signifikant forskjellig fra en forhåndsdefinert verdi (JMP, u.å.). Testen undersøker om studenter har en gjennomsnittlig tillit over den hypotetiske verdien 4, ($\bar{Y} > 4$). Følgende hypotese undersøkes:

H8: Studenter har i gjennomsnitt tillit til at en kunstig intelligens kan være rettferdig i en evalueringsprosess

t-verdi	p-verdi	Cohens d
9,379	< ,0001	0,612

Tabell 4: Resultater fra one-sample t-test

Note: antall frihetsgrader (DF) = 235, Upper 95% mean = 5,3, lower 95% mean = 4,8

For å måle studentenes gjennomsnittlige tillit til kunstig intelligens i en vurderingsprosess, inkluderte spørreundersøkelsen ett spørsmål om studentenes overordnede tillit til problemstillingen, som ble vurdert på en skala fra 1-10. Gjennomsnittet (5,08) gir en indikasjon på det samlede tillitsnivå blant studentene, mens standardavviket (1,76) og variansen (3,08) reflekterer en betydelig variasjon i studentens individuelle tillitsnivå.

Hypotesen undersøkes gjennom en one-sample t-test, der det faktiske gjennomsnittet ble sammenlignet med et hypotetisk gjennomsnitt på 4. Resultatene av t-testen viser en statistisk signifikant sammenheng [t (234) =9,379, p<0,001, d=0,612]. T-verdien på 9,379 og p-verdien på mindre enn 0,001, indikerer at forskjellen mellom det faktiske gjennomsnittet og det hypotetiske gjennomsnittet er svært signifikant. Effektstørrelsen, Cohens d, på 0,612 indikerer en moderat positiv effektstørrelse, noe som betyr at forskjellen mellom gjennomsnittene er betydelig. Analysen viser at det er 95% sikkerhet for at den sanne populasjonsparameteren μ for tillit blant studenter, ligger mellom [4,8, 5,3]. Resultatene gir støtte til hypotesen om at studenter har tillit til kunstig intelligens kan være rettferdig i en evalueringsprosess.

3.2.3 Kji-kvadrattest og krysstabulering

Krysstabulering er en analyseteknikk som anvendes for å undersøke to ulike variabler samtidig (Silkoset et al., 2022, s. 261). Tabellen gir en visuell fremstilling av frekvensfordelingen blant de ulike kategoriene i de to variablene. Dette gjør det mulig å identifisere mønstre og sammenhenger mellom kategoriene. Videre benyttes en kji-kvadrattest for å undersøke om det er en statistisk signifikant sammenheng mellom variablene i tabellen (Silkoset et al., 2022, s. 324). Følgende hypotese testes:

H9: Det er en sammenheng mellom holdningen til eliminering av algoritme aversjon og tillit

	Liten tillit	**Middels tillit**	**Stor tillit**	**Totalt**
Enig i at eliminering av algoritme aversjon øker tillit	38 (12,54%)	162 (53,47%)	18 (5,94%)	221 (72,94)
Uenig i at eliminering av algoritme aversjon øker tillit	2 (0,66%)	1 (0,33%)	2 (0,66%)	5 (1,65%)
Verken eller til at eliminering av algoritme aversjon øker tillit	1 (0,33%)	8 (2,64%)	3 (0,99%)	13 (4,29%)
Totalt	41 (13,53%)	171 (56,44%)	23 (7,59%)	303

Tabell 5: Krysstabell

Antall observasjoner (% av totalen)

Note: Kji^2-verdi 295,94. $p < 0,0001$

Rapportering fra kji-kvadrat i krysstabellen:

Krysstabellen viser at 162 (53,47%) respondenter er enig i at eliminering av algoritme aversjon øker tillit, og opplever en middels tillit til kunstig intelligente systemer. Dermed har flertallet av respondentene denne holdningen. Analysen viser at kun 2 (0,66%) respondenter har en negativ holdning til at eliminering av algoritme øker tillit, og opplever liten tillit til systemet.

Kji-kvadratet på 296 er signifikant på 99,99% nivå (p <,0001). Dette støtter hypotesen som påstår at det er en sammenheng mellom holdningen til eliminering av algoritme aversjon og tillit, og indikerer at det er minst 99,99% sikkerhet for at sammenhengen ikke skyldes tilfeldigheter.

Test	**ChiSquare**	**Prob>ChiSq**
Sannsynlighetsratio	293,594	<,0001
Pearson	295,938	<,0001

Tabell 6: Resultater fra kji-kvadrattest

3.3 Oppsummering av dybdeintervjuene

Tabellen nedenfor presenterer en kortfattet oversikt over de viktigste funnene fra dybdeintervjuene. Funnene er organisert etter sentrale temaer og teorier som er gjennomgående for hele oppgaven. Tabellen gir en oppsummering av de viktigste perspektivene og holdningene som ble observert.

Viktige funn fra dybdeintervjuene:	
Fasitbaserte og reflekterende fag	En konsensus blant intervjuobjektene er at det eksisterer en høyere grad av tillit til at kunstig intelligens kan være rettferdig i en evalueringsprosess knyttet til fasitbaserte fag, sammenlignet med reflekterende fag. Samtlige intervjuobjekter understreker imidlertid at de ikke har tillit til at en kunstig intelligens kan utføre evalueringen alene, selv i fasitbaserte fag. Deres tillit begrenses derfor til å omhandle bruken av kunstig intelligens som en veileder, eller et hjelpemiddel, i evalueringsprosessen.
Bias	Flere av intervjuobjektene uttrykte skepsis og bekymring for bias, og konsekvensene dette kan medføre i form av feilvurderinger og diskriminering. Bias er en utfordring med hensyn til å bygge tillit. Intervjuobjektene understreker behovet for strenge tiltak og systemer som begrenser bias i størst mulig

	grad. Transparente algoritmer, regelmessige evalueringer av systemer og muligheten for menneskelig inngripen utpekes som nødvendig for å overgå bias.
Black box-problemet	En fellesnevner i intervjuene er usikkerheten knyttet til å ha tillit til kunstig intelligens, når man ikke har innsikt i hvilke faktorer som blir vurdert for å ta beslutninger. Intervjuobjektene uttrykker at en vurderingsprosess er svært sensitiv, ettersom feilvurderinger kan medføre store konsekvenser for studentene. Den manglende klarheten rundt hvilke elementer som tas i betraktning i evalueringsprosessen, skaper usikkerhet. Vurderinger og karakterer er av stor betydning for studenter, og det er spesielt ønskelig å vite hvordan systemet har kommet frem til beslutningen, og hvilke elementer som er vektlagt.
Algoritme aversjon	Intervjuobjektenes holdning til å bruke algoritmer var delt. Noen uttrykte skepsis for å stole på algoritmer og uttrykte frykt for å benytte dem i en viktig prosess, som evaluering i utdanning. Disse personene ga uttrykk for usikkerhet til å la en maskin ta beslutninger på deres vegne, og hevder at de har mer tillit til menneskelige beslutninger. Andre hadde en mer positiv og åpen holdning, og så potensialet for å forbedre effektivitet ved bruk av algoritmer. En av intervjuobjektene antydet at eldre personer sannsynligvis har en mer negativ holdning mot bruk av algoritmer sammenlignet med yngre personer.
Tillit	En fellesnevner hos intervjuobjektene er deres moderate tillit til å bruke kunstig intelligens i en vurderingsprosess. Utfordringen med å gi fullstendig tillit til en maskin/robot fremheves. Vurdering av eksamen er et sensitivt tema og det kan oppleves svært urettferdig dersom systemet ikke er pålitelig. På oppfølgingsspørsmål om graden av tillit endres dersom

	systemet kun skal brukes som en veileder, eller at noen kontrollerer svarene, er samtlige positive i større grad.

Tabell 7: Oppsummering av dybdeintervjuene

4.0 Drøfting

Målet med denne studien har vært å undersøke problemstillingen *i hvilken grad har studenter tillit til at en kunstig intelligens kan være rettferdig i en evalueringsprosess*. I denne delen av oppgaven drøftes resultater og funn i sammenheng med problemstillingen. Oppgaven er gjennomgående strukturert i tråd med den teoretiske forankringen, slik at delproblemer og hypoteser er delt inn i kategorier som omfatter fasitbaserte og reflekterende fag, problematikken rundt bias, utfordringene knyttet til black box-modellen og algoritme aversjon. Drøftingen vil derfor også følge denne inndelingen, og behandle hvert av disse punktene individuelt.

4.1 Drøfting av hvert delproblem

Alder:

Hypotese 1 hevdet at alder har en negativ effekt på tillit. Denne hypotesen ble undersøkt gjennom en regresjonsanalyse, men resultatene ga ikke støtte til hypotesen. Analysen viser at p-verdien (0,18) ikke er signifikant på 5% signifikantnivå. Dette indikerer at det ikke er tilstrekkelig grunnlag for å si at det er en signifikant sammenheng mellom variablene. Det er viktig å merke seg at manglende signifikans på p-verdien ikke nødvendigvis betyr at det ikke er noen sammenheng mellom alder og tillit, men heller at det ikke er tilstrekkelig bevis i dataene for å bekrefte sammenhengen. Manglende signifikans på p-verdien kan skyldes ulike faktorer, som for eksempel at utvalgsstørrelsen er for liten, høy variabilitet i dataene eller at det ikke eksisterer en sammenheng mellom variablene.

Fasitbaserte- og reflekterende fag:

Hypotese 2 påsto at fasitbaserte fag har en positiv effekt på tillit, og hypotese 7 påsto at gjennomsnittlig tillit er høyere i fasitbaserte fag sammenlignet med reflekterende fag. Begge hypotesene er basert på tidligere forskning om at kunstig intelligens har en høyere treffsikkerhet i fasitbaserte fag, sammenlignet med reflekterende fag (González-Calatayud et al., 2021). Teknologiske begrensninger gjør at systemet ikke

er optimalt egnet for å evaluere lengre, skriftlige oppgaver. Av den grunn ble det etablert et tydelig skille mellom de to fagområdene, for å identifisere eventuelle ulikheter i graden av tillit. Reflekterende fag som for eksempel psykologi og HR, krever i større grad en subjektiv vurdering fra et menneske. Spesielt for reflekterende fag, kan en potensiell bias være at systemet favoriserer besvarelser som inkluderer spesifikke begreper og teorier, og «straffer» studenter som ikke har inkludert disse elementene. En slik urettferdig bias oppstår som et resultat av at systemet ikke er tilstrekkelig trent på et stort nok treningsdata. På denne måten kan studenter få en urettferdig karakter, til tross for at de faktisk har en tilfredsstillende besvarelse. Selv om også fasitbaserte fag som matematikk og fysikk bør vurderes på skjønn, viser forskning at det er større sannsynlighet for at en kunstig intelligens kan gi en rettferdig evaluering når det er en gitt fasit. Det er likevel viktig å merke seg at begge fagområdene krever en objektiv vurdering for å sikre rettferdighet, og at dette skillet er foretatt for å undersøke et spesifikt felt i forskningen.

Hypotese 2 ble analysert ved hjelp av en regresjonsanalyse, og resultatene viser en signifikant positiv sammenheng [0,18stdBeta, $p < 0,01$], som gir støtte til hypotesen. Dermed kan det konkluderes med at det er en tydelig positiv sammenheng mellom den uavhengige variabelen fasitbaserte fag og den avhengige variabelen tillit. Hypotese 7 ble undersøkt gjennom en independent t-test, og resultatene viser en signifikant sammenheng [$t (255) = 18,49$, $p<0,0001$, $d=1,16$] som gir støtte til hypotesen. Den høye t-verdien (18,49) indikerer en betydelig forskjell i gjennomsnittlig tillit mellom variablene, og den ekstremt lave p-verdien (<,0001) indikerer 99,99% sikkerhet for at resultatene ikke er tilfeldige, og at det er en signifikant sammenheng mellom variablene. Dermed støttes hypotesen. Dybdeintervjuene avdekket også at tilliten til at kunstig intelligens kan være rettferdig i en evalueringsprosess, er høyere innenfor fasitbaserte fag, sammenlignet med reflekterende fag. Her ble det påpekt at intervjuobjektene for øyeblikket ikke har tillit til at systemet kan utføre vurderingen alene, innenfor noen av fagområdene, men at dette kan endres i fremtiden. Tilliten de har, er knyttet til at systemet kan anvendes som en veileder sammen med et menneske. Basert på teori, resultater og funn kan det konkluderes med at studenter har større tillit til at kunstig intelligens kan være

rettferdig i en vurderingsprosess tilknyttet fasitbaserte fag, sammenlignet med reflekterende fag. Det er imidlertid viktig å merke seg at det er grunnleggende begrensninger knyttet til systemets evne til å utføre vurderingen uten menneskelig involvering, noe som gjør det utfordrende å ha full tillit til systemet. Tilliten kan styrkes dersom teknologien utvikles slik at systemet fungerer optimalt uten menneskelig involvering, og oppnår en høyere treffsikkerhet også i reflekterende fag.

Bias:

Hypotese 3 hevdet at bias har en negativ effekt på tillit. Denne hypotesen er basert på tidligere forskning som indikerer at skadelig bias kan resultere i urettferdig eller diskriminerende behandling (Hall & Ellis, 2023). Bias i kunstig intelligens refererer til situasjoner hvor det tas beslutninger som strider med etiske retningslinjer. Baker og Hawn (2022) identifiserer etnisitet, kjønn og nasjonalitet som de mest hyppige formene for bias i utdanningssektoren. Som en konsekvens av bias kan studenter føle usikkerhet knyttet til påliteligheten i det kunstig intelligente systemet. Dermed kan skadelig bias svekke tilliten til systemet.

For å undersøke sammenhengen mellom bias og tillit, ble det gjennomført en regresjonsanalyse. Resultatene fra analysen bekrefter hypotesen ved å vise en statistisk signifikant negativ effekt [-0,16stdBeta, $p < 0,05$]. Dermed kan det konkluderes med at den uavhengige variabelen bias har en negativ effekt på den avhengige variabelen tillit. Dybdeintervjuene identifiserte en betydelig grad av usikkerheten knyttet til at kunstig intelligens kan være preget av bias. Det ble uttrykt bekymring for denne problematikken, og behovet for strenge tiltak og systemer for å begrense bias. I oppgavens teoretiske rammeverk ble det presentert flere tiltak som anbefales å implementeres, inkludert bruk av et representativt treningsdata, reguleringer og etiske retningslinjer, transparens i algoritmene og menneskelig vurdering av algoritmenes resultat (Baker & Hawn, 2022, s. 1052-2092). Implementering av disse tiltakene kan bidra til å redusere bias og øke tilliten til systemet. Funn fra dybdeintervjuene viser at slike tiltak, bidrar til å skape en mer rettferdig og pålitelig evalueringsprosess som studentene kan ha tillit til. Videre forskning og utvikling bør fokusere på å redusere bias før kunstig intelligente systemer tas i bruk i et slik sensitiv domene som utdanningssektoren. Basert på disse

funnene kan det konkluderes med at bias utgjør en utfordring når det gjelder å bygge tillit til kunstig intelligens i en evalueringsprosess, og at det er nødvendig å implementere tiltak for å adressere dette problemet.

Black box-problemet:

Hypotese 4 påsto at black box-problemet har en negativ effekt på tillit. Formålet med hypotesen var å undersøke problematikken med at kunstig intelligente systemer ikke kan gi en forklaring på hvordan de har kommet frem til beslutninger. Selv om det er kjent hvilke data og informasjon som er gitt, og hva som er resultatet, forblir selve beslutningsprosessen ukjent (Gràcia & Sancho-Gil, 2021). Som en konsekvens av dette kan systemet unnlate betydningsfull informasjon som er nødvendig for å fatte rettferdige beslutninger. I en vurderingsprosess kan dette være svært problematisk, da det ikke er mulig å identifisere hvilke faktorer som er inkludert i evalueringen. Dermed begrenses også muligheten til å få en begrunnelse for hvilke elementer som er vektlagt og ikke, for å komme frem til den gitte vurderingen. Black box-modellen gjør det også utfordrende å identifisere bias som påvirker beslutningene, og gjør det dermed utfordrende å avdekke og håndtere slike skjevheter (Gillani et al., 2023, s. 99-111). Som et resultat av dette kan skadelige bias opprettholdes i systemet, og medføre negative konsekvenser for brukerne slik at det er vanskelig å ha tillit til systemet.

Hypotesen undersøkes gjennom en regresjonsanalyse, og resultatene viser en signifikant negativ sammenheng [-0,31stdBeta, $p < 0,001$] mellom de to variablene. Den ekstremt lave p-verdien (< 0,001) viser at sammenhenger en signifikant på 99,9% nivå. Dette gir sterk støtte til hypotesen. Den negative effekten støttes også av funnene fra dybdeintervjuene, som avdekket betydelige utfordringer knyttet til manglende innsikt i algoritmens beslutningsprosess, også kjent som black box-problemet. Dette samsvarer med teorien som er presentert i oppgavens teoretiske del. En konsekvens av black box-problemet er at studenter kan blir tildelt urettferdige karakterer, uten å ha innsikt i hvordan systemet har kommet frem til beslutningen, og om vurdering faktisk er pålitelig. Disse funnene avslører alvorlige utfordringer knyttet til black box-problemet, som utgjør vesentlige hindringer for å etablere tillit til systemet. Økt innsikt i beslutningsprosessen, også kjent som transparens, bidrar til å redusere mistillit og øke studentenes følelse av kontroll. For å adressere black box-

problemet i en vurderingsprosess, er det nødvendig å fremheve ansvarlighet overfor beslutningene. Dermed kreves tydelige retningslinjer og reguleringer for å sikre ansvarlig og pålitelig bruk av systemet, samt å håndtere utfordringene knyttet til black box-modellen.

Eliminering av black box-problemet:

Hypotese 5 hevdet at en hypotetisk eliminering av black box-problemet ville ha en positiv effekt på tillit. Hypotesen er utviklet med utgangspunkt i problematikken knyttet til mangelen av forklarbarhet i beslutningsprosessen til kunstig intelligente systemer (Gràcia & Sancho-Gil, 2021). En hypotetisk eliminering av black box-problemet innebærer å iverksette tiltak som bidrar til å løse problemet. I en studie av Lipton (2017) ble behovet for økt transparens i algoritmene identifisert, med hensikt å øke modellens forklarbarhet og oppdage og korrigere eventuelle bias. I tillegg ble det identifisert et behov for økt innsats innen forskning, og etiske retningslinjer og reguleringer. Videre er det viktig å øke kunnskapen og bevisstheten rundt black box-problemet blant interessenter, for å skape større bevissthet rundt problematikken. Et annet viktig aspekt er å kreve ansvarlighet, slik at det opprettes tydelige ansvarsforhold og strengere krav. Det er viktig å merke seg at denne hypotesen er av spekulativ karakter, og at det ikke er mulig å fjerne utfordringene knyttet til black box-modellen uten videre. Hypotesen er formulert med hensikt å undersøke en teoretisk mulighet for å oppnå eventuelle positive effekter til tillit. Til tross for at den praktiske gjennomførbarheten er begrenset, ønsker analysen av denne hypotesen å bidra til å forstå og utforske potensielle virkninger.

Hypotesen undersøkes i en regresjonsanalyse, og resultatene viser en signifikant sammenheng mellom variablene [0,15stdBeta, $p < 0,01$]. Analysen gir støtte til hypotesen, og det konkluderes med at det foreligger en signifikant, positiv sammenheng mellom variablene. Det understrekes at denne hypotesen representerer et teoretisk scenario, som ikke er realistisk innenfor dagens teknologiske utvikling. Konklusjonen tolkes derfor med forsiktighet, og betraktes som et teoretisk bidrag som belyser potensielle positive effekter dersom en slik løsning oppstår i fremtiden.

En positiv sammenheng mellom eliminering av black box-problemet og tillit, kan ha implikasjoner som strekker seg utover den spesifikke konteksten. Denne sammenhengen kan overføres til andre domener, for eksempel helsevesenet, finans eller autonome kjøretøy. Når black box-problemet håndteres og tilliten øker, blir det mer sannsynlig at samfunnet vil akseptere bruken av kunstig intelligens i flere områder. I tillegg har økt transparens, åpne beslutningsprosesser og ansvarlighet en synergisk effekt for å styrke tilliten, både med tanke på black box-problemet, bias og algoritme aversjon. Tydelige retningslinjer og reguleringer som sikrer etisk og forsvarlig bruk av kunstig intelligente systemer bidrar til å bygge tillit hos brukerne. Ved å ha en mer transparent og ansvarlig tilnærming til bruken av kunstig intelligens, kan man motvirke bias og sikre en rettferdig og pålitelig modell. Samlet sett bidrar slike tiltak, til økt tillit og en styrket aksept for bruk av kunstig intelligens i ulike sektorer. Dette legger grunnlaget for en mer effektiv og etisk bruk av teknologien, og åpner opp for potensielle fordeler og innovasjoner i samfunnet.

Eliminering av algoritme aversjon:

Hypotese 6 innebærer at en hypotetisk eliminering av algoritme aversjon vil ha en positiv effekt på tillit, og hypotese 9 påsto at det er en sammenheng mellom holdningen til eliminering av algoritme aversjon og tillit. Begge hypotesene er formulert i samsvar med teorier som antyder at mange mennesker er skeptiske til å bruke algoritmer, til tross for at forskning viser at algoritmer konsekvent er mer treffsikre enn menneskelige beslutninger (Dietvorst et al., 2018, s. 1155-1170). Det er verdt å merke seg at begge hypotesene er av spekulativ art, og har som formål å undersøke om en hypotetisk eliminering kunne medført økt tillit. Ifølge Dietvorst et al. (2015) har mennesker en tendens til å ha større tillit til at mennesker kan ta rettferdige beslutninger, i forhold til kunstig intelligente systemer. Tidligere forskning viser også at folk er mer åpne for å anvende algoritmer dersom de har muligheten til å kontrollere algoritmens prognose (Dietvorst et al., 2018, s. 1155-1169). Algoritme aversjon er en subjektiv holdning som ikke kan fjernes uten videre. Derfor representerer disse hypotesene kun en teoretisk konstruksjon av en situasjon som for øyeblikket ikke er realistisk, og det er lite sannsynlig at den blir realisert i nær fremtid. Formålet med hypotesene er å undersøke om studenter kan oppleve økt tillit

til kunstig intelligens i en evalueringsprosess, dersom algoritme aversjon elimineres. Dette hypotetiske scenarioet, betyr i praksis at det ikke vil eksistere en skepsis mot algoritmer.

Hypotese 6 ble undersøkt gjennom en regresjonsanalyse, og resultatene avdekker at det ikke er en signifikant sammenheng mellom eliminering av algoritme aversjon og tillit [0,02, p = NS]. Dermed forkastes hypotesen. Regresjonsanalysen undersøker om det er en lineær sammenheng mellom variablene, så det kan likevel eksistere en sammenheng mellom variablene, men at sammenhengen ikke er lineær. Analysen viser en standardisert regresjonsparameter 0,01, noe som indikerer at sammenhengen er positiv, men at den ikke er lineær ifølge p-verdien. Hypotese 9 ble undersøkt ved hjelp av krysstabulering og en kji-kvadrattest. Analysen av krysstabellen viser at 162 (53,47%) respondenter er enig i at eliminering av algoritme aversjon øker tillit, og opplever en middels tillit til kunstig intelligente systemer i en vurderingsprosess. Dette indikerer at flertallet av det totale antallet observasjoner har denne holdningen. Resultatene fra kji-kvadrattesten identifiserer at sammenhengen er signifikant. Kji-kvadratet på 296 er signifikant på 99,99% nivå. Dette gir støtte for hypotesen og det kan konkluderes med at det er en sammenheng mellom holdningen til eliminering av algoritme aversjon og tillit, med 99,99% sikkerhet. Funn fra dybdeintervjuene viser at holdningen til algoritmer er delt. Noen uttrykker at de har mer tillit til menneskelige beslutninger fremfor algoritmer, mens andre sier det motsatte.

Selv om det ikke er en lineær sammenheng mellom variablene ifølge regresjonsanalysen, kan det fremdeles være en signifikant sammenheng på et mer overordnet nivå. Selv om hypotese 6 ble forkastet i regresjonsanalyse, ble hypotese 9 beholdt i kji-kvadrattesten, og kan derfor indikere at det eksisterer en sammenheng mellom variablene, men at sammenhengen ikke er lineær. Regresjonsanalysen undersøker om det er en lineær sammenheng mellom variablene og tar hensyn til andre faktorer samtidig, mens kji-kvadrattesten undersøker om det er en mulig assosiasjon mellom variablene, uavhengig av andre faktorer. Dette betyr at selv om det ikke er en lineær sammenheng mellom eliminering av algoritme aversjon og tillit, kan det likevel være en betydningsfull positiv sammenheng. Det konkluderes derfor

med at en hypotetisk eliminering av algoritme aversjon har en sammenheng med tillit, men at det er usikkerhet hvor tydelig sammenhengen er.

Tillit:

Formålet med hypotese 8 var å måle studentenes generelle tillit til at en kunstig intelligens kan være rettferdig i en evalueringsprosess. Hypotesen ble formulert for å undersøke studentenes overordnende tillit, i samsvar med problemstillingen. Tillit til en kunstig intelligens handler om å ha tillit til at systemet har den nødvendige teknologien som kreves for å oppfylle de gitte formålene. Mennesker har ulik oppfatning av tillit, men tillit er avgjørende for å ta i bruk teknologien (Qin et al., 2020, s.1693-1710). Brukerne må kunne stole på at systemet har egenskapene til å fatte rettferdige beslutninger, og beskytte personvernet. Derfor er det viktig å påpeke at et kunstig intelligent system ikke kan tilskrives menneskelige egenskaper, og har hverken selvbevissthet eller evnen til å ta egne valg. Dermed er det avgjørende å vektlegge tiltak og retningslinjer som bidrar til å styrke systemets pålitelighet, for å bygge tillit blant studentene.

For å teste gjennomsnittet i utvalget ble det benyttet en one-sample t-test. Testen sammenlignet studentenes tillit mot en hypotetisk gjennomsnittsverdi ($\bar{Y} > 4$). Resultatene av t-testen avslørte en signifikant sammenheng [t (234) =9,379, p < 0,001, d =0,612], og gir derfor støtte til hypotesen. Konklusjonen er at studentene i gjennomsnitt har tillit til at en kunstig intelligens kan være rettferdig i en evalueringsprosess. Dette antyder at studentene har en viss aksept og positiv holdning til bruken av systemet.

Selv om t-testen indikerer at studentene opplever tillit, er det viktig å erkjenne at tillit er en flytende faktor som krever kontinuerlig innsats for å opprettholdes. Tillit kan raskt brytes ved feil valg, og er vanskelig å bygge opp igjen. Derfor er det avgjørende å implementere nødvendige tiltak for å sikre at systemet opprettholder sin integritet og leverer pålitelige og rettferdige vurderinger. Mayer, Davis og Schoorman (1995) har utviklet en av de mest anerkjente modellene for å forstå tillit. Modellen består av tre grunnleggende komponenter som påvirker tillit: kompetanse, integritet og velvilje. Kompetanse handler om at systemet har nødvendig teknologi, kapasitet og ferdigheter

til å utføre pålitelige beslutninger. Integritet innebærer at systemet opptrer på en ærlig, pålitelig og rettferdig måte. Det handler om at systemet opererer i tråd med etiske retningslinjer og prinsipper. Kunstig intelligente systemer har integritet dersom det tar hensyn til transparens, rettferdighet og personvern. Velvilje handler om at systemet tar beslutninger som er til fordel for brukeren. Tillit til kunstig intelligens innebærer å ha tillit til at systemet er designet med en intensjon om å skape verdi og nytte i brukerens favør. Kompetanse, integritet og velvilje er gjensidig avhengig av hverandre og gir en synergisk effekt når det gjelder å påvirke tilliten til kunstig intelligens. For å bygge og opprettholde tillit til kunstig intelligens i en vurderingsprosess, er det viktig å legge disse prinsippene til grunn. Ved å fokusere på disse komponentene, bidrar man til at brukerne føler seg trygge på systemets evne til å fatte rettferdige vurderinger, slik at tillit kan økes. Dette er avgjørende for å sikre en vellykket implementering av kunstig intelligens i vurderingsprosesser, og oppnå aksept og tillit fra studentene.

4.2 Svakheter ved studien og anbefalinger for fremtidig forskning

Kunstig intelligens er ett fagfelt i stadig utvikling. Teknologien utvikler seg raskere enn forskningen, og det har derfor vært utfordrende å finne relevant empirisk materiale som er oppdatert den siste tiden. Dette kan ha begrensninger når det gjelder studiens tilgang på oppdatert informasjon og innsikt i den nyeste utviklingen.

En annen begrensning er at utvalget i studien begrenser seg til Norge. Kunstig intelligens i utdanning er mer utbredt i andre land. I for eksempel USA og England, har det allerede blitt gjennomført flere prosjekter med kunstig intelligens innenfor utdanning. Dette betyr at funnene i studien kanskje ikke kan generaliseres til andre land, hvor mye av den aktuelle forskningen blir utført. For fremtidig forskning anbefales det derfor å undersøke problemstillingen også i andre land hvor fenomenet er mer utbredt.

En ytterligere svakhet er at det kan være utfordrende for respondentene å forstå fenomenet knyttet til kunstig intelligens. Fagområdet er komplekst, og det kan være vanskelig å forstå hvordan det fungerer i praksis. Variablene og delproblemene i studien bygger på relevant teori og forskning, noe som kan gjøre det vanskelig for respondentene å forstå hva dette betyr uten å ha dypere kunnskap om emnet. For

eksempel, kan begrepet algoritme aversjon virke urealistisk eller uklart for respondentene. Dette kan muligens forklare hvorfor hypotese 8 ikke viste seg å være signifikant i regresjonsanalysen. Undersøkelsen ble gjennomført utenfor kontrollerte forhold, noe som kan føre til at deltakernes svar inneholder subjektive oppfatninger. Ved hvert spørsmål i spørreundersøkelsen ble det presentert relevant teori knyttet til hvert delproblem som ble undersøkt. Likevel er det en sannsynlighet for at respondentene kan ha mistolket spørsmålene og dermed gitt feilaktig svar. Dette kan skape bias i resultatene og gjøre det utfordrende å validere om deltakernes svar faktisk viser deres ærlige holdninger. Disse begrensningen utgjør en trussel mot studiens validitet. Av den grunn anbefales det å gjennomføre test-retest for å kunne vurdere reliabiliteten i videre forskning.

En annen potensiell svakhet ved studien er at relevante variabler som kan ha sammenheng med tilliten er ekskludert. Dette kan medføre negative implikasjoner for forskningens validitet og generaliserbarhet. Det er derfor nødvendig å være oppmerksom på begrensningene forbundet med eventuelle utelatte variabler og deres mulige innvirkning på studiens resultater. Til videre forskning foreslås det å inkludere en variabel knyttet til studentenes faktiske kunnskap om fenomenet. En slik variabel kunne bidratt til å øke forklaringskraften til regresjonsmodellen.

5.0 Konklusjon

Det overordnede formålet med denne oppgaven har vært å besvare problemstillingen: *i hvilken grad har studenter tillit til at en kunstig intelligens kan være rettferdig i en evalueringsprosess*. Resultater og funn avslører at studenter har en moderat tillit til fenomenet. Studien avslører at studenter har en større grad av tillit i fasitbaserte fag, sammenlignet med reflekterende fag. Samtidig er det identifisert flere utfordringer knyttet til å ha tillit til kunstig intelligens. Bias, black box-problemet og algoritme aversjon er faktorer som preger studentenes tillit til systemet. Imidlertid viser funn fra studien at studentenes tillit kan økes dersom disse elementene elimineres. Dette innebærer blant annet å sikre et bredere mangfold i treningsdataene, slik at bias kan reduseres. Det anbefales også å sikre økt transparens i algoritmene, slik at systemets forklarbarhet økes. På denne måten oppnås mer innsikt i systemets beslutningsprosesser, og bidrar til å løse black box-problemet. I tillegg anbefales det å

innføre tiltak som sørger for pålitelig bruk av systemene, og ansvarlighet overfor beslutningene som tas.

I denne oppgaven er det avdekket funn som gir grunnlag for å konkludere med at studenter har en moderat tillit til at kunstig intelligens kan være rettferdig i en evalueringsprosess. Funnene indikerer at studentenes tillit kan økes ytterligere dersom det iverksettes tiltak for å redusere bias og algoritme aversjon, og løse black box-problemet.

6.0 Referanseliste

Aloisi, C. (2023). The future of standardised assessment: Validity and trust in algorithms for assessment and scoring. *European Journal of Education*, 58, 98– 110. https://doi-org.ezproxy.library.bi.no/10.1111/ejed.12542

Arnold, M. H. (2021). Teasing out Artificial Intelligence in Medicine: An Ethical Critique of Artificial Intelligence and Machine Learning in Medicine. Journal of Bioethical Inquiry, 18, 121-139. https://doi.org/10.1007/s11673-020-10080-1

Asaro, P. M. (2019). AI ethics in predictive policing: From models of threat to an ethics of care. *IEEE Technology and Society Magazine, 38*(2), 40–53. https://doi.org/10.1109/MTS.2019.2915154.

Ashish, J., O. (u.å.) Grappling with the prospect of AI in education. *Questmite*. Hentet 29.05.23 fra https://questmite.com/technology/grappling-with-the-prospect-of-ai-in-education/

Aslam, F., Hunjra, A. I., Ftiti, Z., Louhichi, W. & Shams, T. (2022) Insurance fraud detection: Evidence from artificial intelligence and machine learning. *Research in International Business and Finance*, 62, 2-9 https://doi.org/10.1016/j.ribaf.2022.101744

Baker, R. S. & Hawn, A. (2022). Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education* 32, 1052–1092 https://doi-org.ezproxy.library.bi.no/10.1007/s40593-021-00285-9

Brown, T. B., Mané, D., Roy, A., Abadi, M. & Gilmer, J. (2018). Adversarial Patch. https://doi.org/10.48550/arXiv.1712.09665

Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research,* 81, 77-91.

Cohen J. (1988). Statistical power analysis for the behavioral sciences. 2. utg. Hillsdale, New Jersey: Lawrence Erlbaum Associates

Conradsen, S. (2022) Tillit. *Store norske leksikon*. Hentet 29.05.23 fra https://snl.no/tillit

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. Hentet 23.05.23 fra https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Datatilsynet (2019) *Anonymisering av personopplysninger* https://www.datatilsynet.no/rettigheter-og-plikter/virksomhetenes-plikter/informasjonssikkerhet-internkontroll/hvordan-anonymisere-personopplysninger/

Datta, A., Tschantz, M. C. & Datta, A. (2015) Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies Symposium.* 1, 92-112 https://doi.org/10.1515/popets-2015-0007

Dietvorst, B. J., Simmons, J. P. & Massey, C. (2015) Algorithm aversion: people erroneously avoid algorithms after seeing them err *Journal of Experimental Psychology: General,* 144(1), s. 114–126 http://dx.doi.org/10.1037/xge0000033.supp

Dietvorst, B. J., Simmons, J. P. & Massey, C. (2018) Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science* 64(3),1155-1170. https://doi.org/10.1287/mnsc.2016.2643

Dvergsdal, H. & Karlsen, G. (2023) Turingtesten. *Store norske leksikon*. Hentet 08.02.23 fra https://snl.no/turingtesten

Einhorn, H. J. (1986). Accepting error to make less error. *Journal of Personality Assessment, 50,* 387–395. http://dx.doi.org/10.1207/s15327752jpa5003_8

Elements of ai. (u.å.) *Hva er kunstig intelligens?* Hentet 06.02.23 fra https://course.elementsofai.com/no/1/3

European Commission (2019, 08.04). *A definition of Artificial Intelligence: main capabilities and scientific disciplines.* Hentet 08.03.23 https://digital-

strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

Gillani, N., Eynon, R., Chiabaut, C., & Finkel, K. (2023). Unpacking the "Black Box" of AI in Education. *Educational Technology & Society, 26*(1), 99-111. https://doi.org/10.30191/ETS.202301_26(1).0008

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660. https://doi-org.ezproxy.library.bi.no/10.5465/annals.2018.0057

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial Intelligence for Student Assessment: A Systematic Review. *Applied Sciences*, *11*(12), 5467 https://doi.org/10.3390/app11125467

Gràcia, X. G. & Sancho-Gil, J. M. (2021). Artificial Intelligence in Education: Big Data, Black Boxes, and Technological Solutionism. *Seminar.Net, 17*(2). https://doi.org/10.7577/seminar.4281

Grønmo, S. (2020). Bias i forskning. I Store norske leksikon. Hentet 23.05.23 fra https://snl.no/bias_i_forskning

Grønmo, S. (2023a). Kvalitativ metode. I *Store norske leksikon*. Hentet 17.04.23 fra https://snl.no/kvalitativ_metode

Grønmo, S. (2023b). Kvantitativ metode. I *Store norske leksikon*. Hentet 17.04.23 fra https://snl.no/kvantitativ_metode

Hall, P. & Ellis, D. (2023*).* A systematic review of socio-technical gender bias in AI algorithms. *Online Information Review* https://doi-org.ezproxy.library.bi.no/10.1108/OIR-08-2021-0452

IBM (u.å.a) *Deep Blue*. IBM. Hentet 08.02.23 fra https://www.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/

IBM. (u.å.b) *IBM Watson.* Hentet 07.02.23 fra https://www.ibm.com/watson

JMP. (u.å.). The One Sample t-Test. Hentet 29.05.23 fra https://www.jmp.com/en_sg/statistics-knowledge-portal/t-test/one-sample-t-test.html

Johannessen, A., Tufte, P., A. & Christoffersen, L. (2011). *Introduksjon til samfunnsvitenskapelig metode.* (4. utg.) Abstrakt forlag

Kahneman, D. (2011*). Thinking, fast and slow*. Macmillan.

Kahneman, D., Rosenfield, A. M., Gandhi, L., & Blaser, T. (2016). Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making. *Harvard Business Review*.

Klašnja-Milićević, A., Ivanović, M. & Budimac, Z. (2017). Data science in education: big data and learning analytics. *Computer Application in Engineering Education*, 25, 1066– 1078. https://doi-org.ezproxy.library.bi.no/10.1002/cae.21844

Konam, S. (2022) *Where did IBM go wrong with Watson Health?* Quartz. Hentet 07.02.23 fra https://qz.com/2129025/where-did-ibm-go-wrong-with-watson-health

Kunnskapsdepartementet (2023). *Strategi for digital kompetanse og infrastruktur i barnehage og skoler*. Hentet 25.04.2023 fra https://www.regjeringen.no/no/dokumenter/strategi-for-digital-kompetanse-og-infrastruktur-i-barnehage-og-skole/id2972254/?ch=1

Lipton, Z. C. (2017). The Mythos of Model Interpretability. https://doi.org/10.48550/arXiv.1606.03490

Mayer, R. C., Davis, J. H. & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.2307/258792

Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence*, 2, 45-47 https://doi.org/10.1016/j.caeai.2021.100020.

Øye, O. J. & Normann, M. (2021) *Dette er utfordringene med kunstig intelligens*. Oslomet. Hentet 09.02.23 fra https://www.oslomet.no/forskning/forskningsnyheter/dette-er-utfordringene-med-kunstig-intelligens

Prahl, A. & Swol, L. V. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, 36(6), 691-702 https://doi-org.ezproxy.library.bi.no/10.1002/for.2464

Qin, L., Kai, L. & Yan, J. (2020). Understanding user trust in artificial intelligence-based educational systems: Evidence from China. *British journal of educational technology* 51(5), 1693-1710. https://doi-org.ezproxy.library.bi.no/10.1111/bjet.12994

Qlik (2023) *Big Data AI.* Hentet 09.02.23 fra https://www.qlik.com/us/augmented-analytics/big-data-ai

Raczynski, K. & Cohen, A. (2018) Appraising the scoring performance of automated essay scoring systems—Some additional considerations: Which essays? Which human raters? Which scores? Applied Measurement in Education, 31(3), 233-240 https://doi-org.ezproxy.library.bi.no/10.1080/08957347.2018.1464449

Ramesh, D. & Sanampudi, S.K. (2022) An automated essay scoring system: a systematic literature review. *Artificial Intelligence Review,* 55, 2495–2527 https://doi-org.ezproxy.library.bi.no/10.1007/s10462-021-10068-2

Reich, T., Kaju, S. & Maglio, S. J. (2022). How to overcome algorithm aversion: Learning from mistakes. *Journal of Consumer Psychology*, 33(2), 285-302. https://doi-org.ezproxy.library.bi.no/10.1002/jcpy.1313

Ringdal, K. (2018) *Enhet og mangfold, samfunnsvitenskapelig forskning og kvantitativ metode* (4. utg.). Fagbokforlaget.

Roll, I. & Wylie, R. (2016). Evolution and Revolution in Artificial Intelligence in Education. International Journal of Artificial Intelligence in Education, 26, 582-599. https://doi.org/10.1007/s40593-016-0110-3

Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. Science and Engineering Ethics, 26, 2749-2767. https://doi-org.ezproxy.library.bi.no/10.1007/s11948-020-00228-y

Schiff, D. (2022). Education for AI, *not* AI for Education: The Role of Education and Ethics in National AI Policy Strategies. *International Journal of Artificial Intelligence in Education*, 32, 527–563 https://doi-org.ezproxy.library.bi.no/10.1007/s40593-021-00270-2

Silkoset, R. Olsson, U. H. & Gripsrud, G. (2021). *Metode, datanalyse og innsikt.* (4.utg). Cappelen damm akademisk

Sjåstad, H. (2019) *Algoritme-aversjon*. Magma, s. 63-70. Hentet 20.02.23 fra https://old.magma.no/algoritme-aversjon

Tidemann, A. (2021a) Ekspertsystem. *Store norske leksikon*. Hentet 08.02.23 fra https://snl.no/ekspertsystem#-Historikk

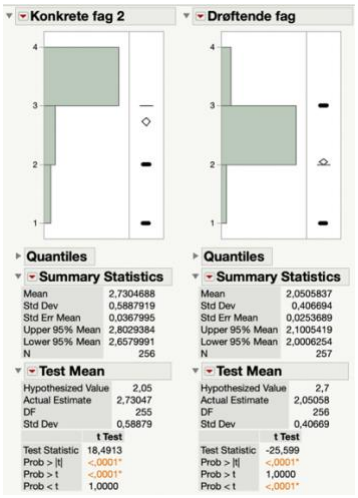
Tidemann, A. (2023b) Kunstig intelligens. *Store norske leksikon*. Hentet 08.02.23 fra https://snl.no/kunstig_intelligens

Tidemann, A. & Elster, A. C., (2022) Maskinlæring. *Store norske leksikon*. Hentet 08.02.23 fra https://snl.no/maskinlæring

Young, S. (2020). The future of farming: Artificial intelligence and agriculture. *Harvard International Review, 41*(1), 45-47.

7.0 Vedlegg

Independent t-test av H7:



Konkrete fag 2

Quantiles

Summary Statistics

Mean	2,7304688
Std Dev	0,5887919
Std Err Mean	0,0367995
Upper 95% Mean	2,8029384
Lower 95% Mean	2,6579991
N	256

Test Mean

Hypothesized Value	2,05
Actual Estimate	2,73047
DF	255
Std Dev	0,58879

	t Test
Test Statistic	18,4913
Prob > |t|	<,0001*
Prob > t	<,0001*
Prob < t	1,0000

Drøftende fag

Quantiles

Summary Statistics

Mean	2,0505837
Std Dev	0,406694
Std Err Mean	0,0253689
Upper 95% Mean	2,1005419
Lower 95% Mean	2,0006254
N	257

Test Mean

Hypothesized Value	2,7
Actual Estimate	2,05058
DF	256
Std Dev	0,40669

	t Test
Test Statistic	-25,599
Prob > |t|	<,0001*
Prob > t	1,0000
Prob < t	<,0001*

One-sample t-test av H8:

Summary Statistics

Mean	5,0765957
Std Dev	1,7597364
Std Err Mean	0,1147925
Upper 95% Mean	5,3027547
Lower 95% Mean	4,8504368
N	235

Test Mean

Hypothesized Value	4
Actual Estimate	5,0766
DF	234
Std Dev	1,75974

	t Test
Test Statistic	9,3786
Prob > |t|	<,0001*
Prob > t	<,0001*
Prob < t	1,0000

Regresjonsanalyse:

Summary of Fit

RSquare	0,153884
RSquare Adj	0,131321
Root Mean Square Error	1,648442
Mean of Response	5,077586
Observations (or Sum Wgts)	232

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	111,19740	18,5329	6,8202
Error	225	611,40605	2,7174	Prob > F
C. Total	231	722,60345		<,0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%	Std Beta
Intercept	5,2082164	1,359423	3,83	0,0002*	2,5293878	7,8870449	0
Alder	-0,138291	0,10413	-1,33	0,1855	-0,343487	0,0669039	-0,08198
Konkrete fag 2	0,5473612	0,201564	2,72	0,0071*	0,1501658	0,9445566	0,178552
Bias 2	-0,368185	0,166813	-2,21	0,0283*	-0,6969	-0,03947	-0,16065
Black box problemet	-0,805245	0,182002	-4,42	<,0001*	-1,163891	-0,446599	-0,312
Black box problemet eksisterer ikke	0,8559749	0,371996	2,30	0,0223*	0,1229333	1,5890165	0,152369
Overkomme algoritme aversjon	0,0251946	0,168101	0,15	0,8810	-0,306059	0,3564483	0,010371

Tillit = 5,2 – 0,14alder + 0,55fasitbaserte fag – 0,37bias – 0,8black box + 0,85xblack box + 0,02algoritme aversjon

Krysstabell og kji-kvadrattest:

Contingency Table

Tillit 4

Algoritme aversjon 3	Count / Total % / Col % / Row %	.	Liten	Middels	Stor	Total
	.	64	0	0	0	64
		21,12	0,00	0,00	0,00	21,12
		94,12	0,00	0,00	0,00	
		100,00	0,00	0,00	0,00	
	Enig	3	38	162	18	221
		0,99	12,54	53,47	5,94	72,94
		4,41	92,68	94,74	78,26	
		1,36	17,19	73,30	8,14	
	Uenig	0	2	1	2	5
		0,00	0,66	0,33	0,66	1,65
		0,00	4,88	0,58	8,70	
		0,00	40,00	20,00	40,00	
	Verken eller	1	1	8	3	13
		0,33	0,33	2,64	0,99	4,29
		1,47	2,44	4,68	13,04	
		7,69	7,69	61,54	23,08	
	Total	68	41	171	23	303
		22,44	13,53	56,44	7,59	

Tests

N	DF	-LogLike	RSquare (U)
303	9	146,79710	0,4308

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	293,594	<,0001*
Pearson	295,938	<,0001*

Cronbachs alfa:

Cronbach's α

	α	-,8 -,6 -,4 -,2 0 ,2 ,4 ,6 ,8
Entire set	0,7134	

Excluded Col	α	-,8 -,6 -,4 -,2 0 ,2 ,4 ,6 ,8
Konkrete fag	0,6963	
Drøftende fag 2	0,6978	
Bias	0,6549	
Overgår bias	0,6922	
Black box	0,6440	
Løse problemet	0,6951	
Algoritme aversjon	0,6748	

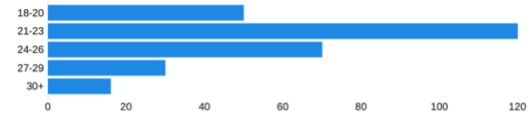
Resultatene fra spørreundersøkelsen:

- Hvilket kjønn er du?

Field	Min	Max	Mean	Standard Deviation	Variance	Responses
Data	1.00	2.00	1.60	0.49	0.24	284

Field		Choice Count
Kvinne	40.14%	114
Mann	59.86%	170
Total		284

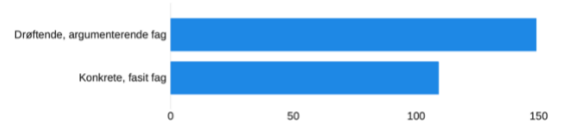
- Hvor gammel er du?



Field	Min	Max	Mean	Standard Deviation	Variance	Responses
Data	1.00	5.00	2.45	1.07	1.14	286

Field		Choice Count
18-20	17.48%	50
21-23	41.96%	120
24-26	24.48%	70
27-29	10.49%	30
30+	5.59%	16
Total		286

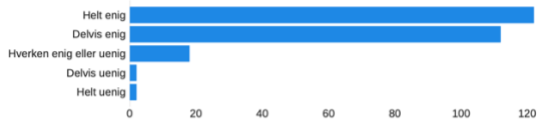
- Vi skiller mellom studieretningen som har fag hvor eksamen går ut på å drøfte og å argumentere (psykologi, HR, filosofi osv) og fag hvor det mer eller mindre er en fasit (matte, økonomi, fysikk osv). Hvilken av disse to studieretningene tilhører du?



Field	Min	Max	Mean	Standard Deviation	Variance	Responses
Data	1.00	2.00	1.42	0.49	0.24	258

Field		Choice Count
Drøftende, argumenterende fag	57.75%	149
Konkrete, fasit fag	42.25%	109
Total		258

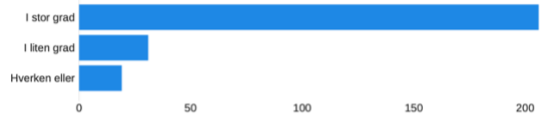
- Er du enig / uenig i at studenters studieretning kan påvirke graden av tillit til at KI retter eksamen?



Field	Min	Max	Mean	Standard Deviation	Variance	Responses
Data	1.00	5.00	1.63	0.72	0.51	256

Field	Choice Count		
Helt enig	47.66%	122
Delvis enig	43.75%	112
Hverken enig eller uenig	7.03%	18
Delvis uenig	0.78%	2
Helt uenig	0.78%	2
Total		256

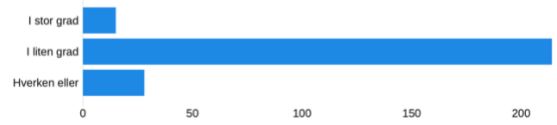
- I hvilken grad har du tillit til at en KI kan gi en rettferdig karakter til fag med en gitt fasit?



Field	Min	Max	Mean	Standard Deviation	Variance	Responses
Data	1.00	3.00	1.27	0.59	0.35	256

Field	Choice Count		
I stor grad	80.47%	206
I liten grad	12.11%	31
Hverken eller	7.42%	19
Total		256

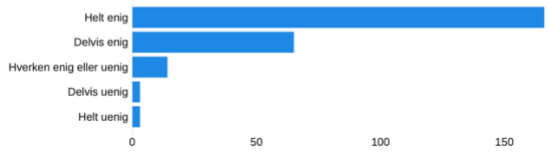
- I hvilken grad har du tillit til at en KI kan gi en rettferdig karakter til en skriftlig, drøftende tekst?



Field	Min	Max	Mean	Standard Deviation	Variance	Responses
Data	1.00	3.00	2.05	0.41	0.16	257

Field	Choice Count		
I stor grad	5.84%	15
I liten grad	83.27%	214
Hverken eller	10.89%	28
Total		257

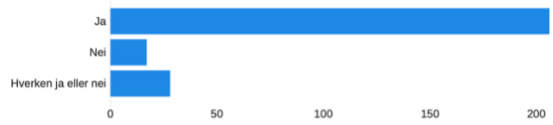
- Bias defineres som systematisk skjevhet og refererer til fordommer eller forhåndsinnstillinger som kan påvirke beslutninger og handlinger. KI-modellen vil være trent på data fra tidligere eksamensoppgaver, og karakteren som en menneskelig sensor har gitt. Dette gjør at dersom du ikke inkluderer ord og uttrykk som tidligere har fått poeng, kan du få en lavere karakter selvom svaret ditt egentlig er bedre. På denne måten kan KI-modellen skape bias, altså skjevheter og diskriminering i dens beslutning. Svekker dette din tillit til at KI kan rette eksamen?



Field	Min	Max	Mean	Standard Deviation	Variance	Responses
Data	1.00	5.00	1.45	0.76	0.57	251

Field	Choice Count
Helt enig	66% 166
Delvis enig	26% 65
Hverken enig eller uenig	6% 14
Delvis uenig	1% 3
Helt uenig	1% 3
Total	251

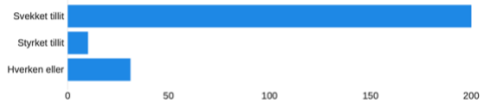
- Dersom du får vite at denne typen bias i stor grad kan unngås dersom KI-modellen trenes på et bredere spekter av data, som ikke favoriseres visse typer svar - vil dette styrke din tillit til at KI retter eksamen?



Field	Min	Max	Mean	Standard Deviation	Variance	Responses
Data	1.00	3.00	1.29	0.66	0.43	251

Field	Choice Count
Ja	82.07% 206
Nei	6.77% 17
Hverken ja eller nei	11.16% 28
Total	251

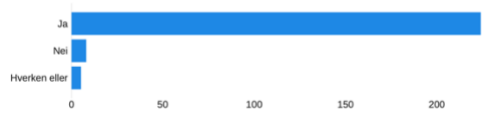
- En utfordring med kunstig intelligens er at algoritmene er så komplekse at det kan være vanskelig å finne ut hvordan modellen har kommet frem til resultatet. Det er kompliserte sammenhenger mellom input i en slik modell, og konklusjonen. Dermed vil ikke modellen kunne gi en begrunnelse for hvilke variabler som er vurdert for å komme frem til karakteren. Hvordan påvirker dette din tillit til at en KI gir en rettferdig karakter?



Field	Min	Max	Mean	Standard Deviation	Variance	Responses
Data	1.00	3.00	1.30	0.68	0.47	241

Field		Choice Count
Svekket tilit	82.99%	200
Styrket tilit	4.15%	10
Hverken eller	12.86%	31
Total		241

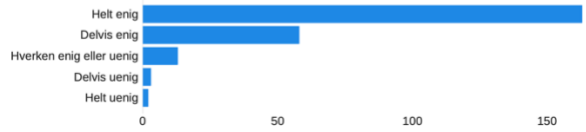
- Ville du hatt mer tillit til KI-modellenen dersom den ga en begrunnelse for hvorfor du fikk den aktuelle karakteren?



Field	Min	Max	Mean	Standard Deviation	Variance	Responses
Data	1.00	3.00	1.08	0.34	0.11	237

Field		Choice Count
Ja	94.51%	224
Nei	3.38%	8
Hverken eller	2.11%	5
Total		237

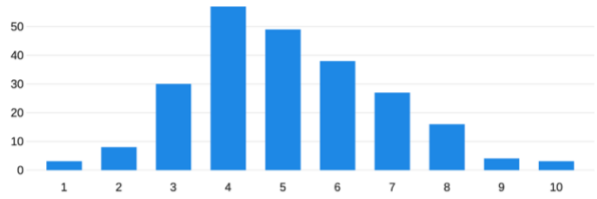
- Algoritme aversjon handler om menneskers skepsis mot å stole på algoritmer. Noen mener at folk vil være mer åpne for å stole på algoritmer dersom de, eller noen andre, får mulighet til å kontrollere algoritmenes prognose. Hvor enig er du i denne påstanden: "Min tillit til at KI retter eksamen styrkes dersom jeg vet at en lærer kontrollerer svarene"



Field	Min	Max	Mean	Standard Deviation	Variance	Responses
Data	1.00	5.00	1.42	0.73	0.53	239

Field	Choice Count	
Helt enig	68.20%	163
Delvis enig	24.27%	58
Hverken enig eller uenig	5.44%	13
Delvis uenig	1.26%	3
Helt uenig	0.84%	2
Total		239

- Basert på din nåværende kunnskap om kunstig intelligens, hvordan vil du beskrive din tillit til at KI anvendes for å rette eksamen på en skala fra 1-10? Hvor 1 er svak, og 10 er sterk.



Field	Min	Max	Mean	Standard Deviation	Variance	Responses
Data	1.00	10.00	5.08	1.76	3.08	235

Field	Choice Count	
1	1%	3
2	3%	8
3	13%	30
4	24%	57
5	21%	49
6	16%	38
7	11%	27
8	7%	16
9	2%	4
10	1%	3
Total		235