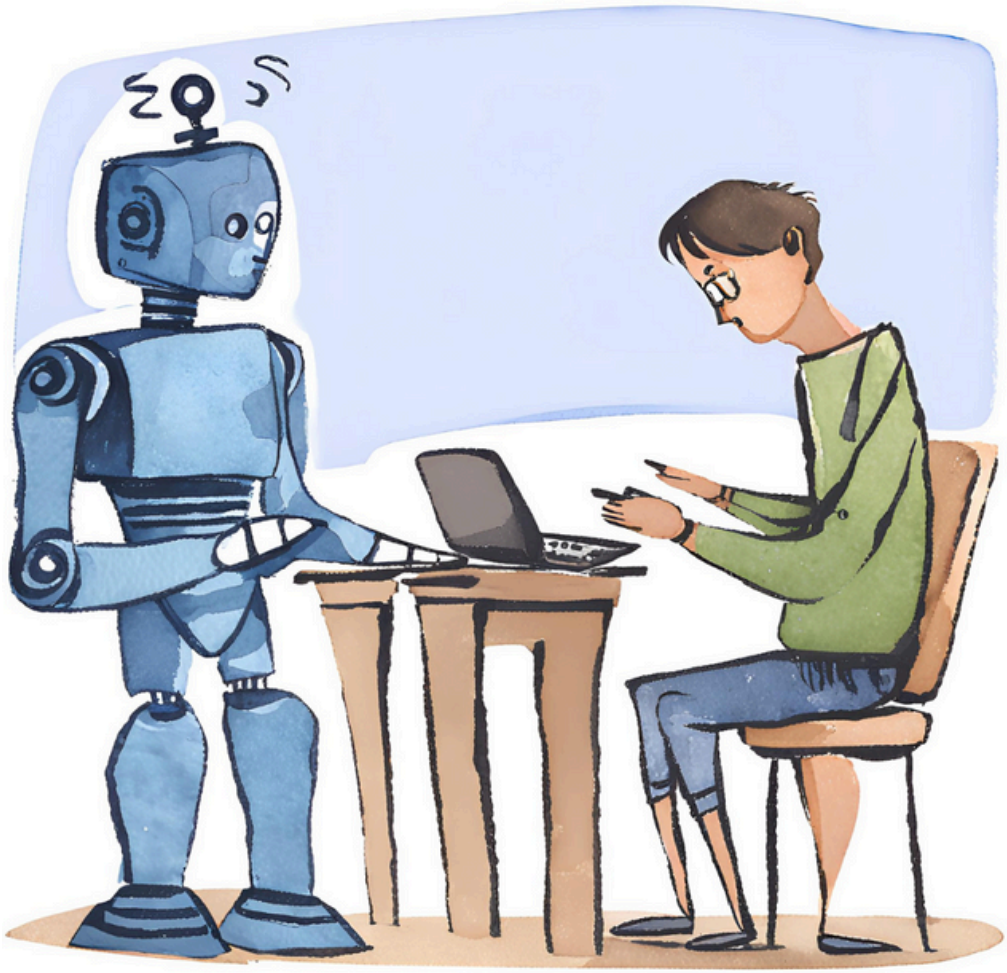
Kan AI blive gymnasielærernes nye kollega?



Et eksplorativt studie om brugen af AI som beslutningsstøtteværktøj til at højne korrektheden af bedømmelser af danskopgaver i gymnasiet

Kan AI blive gymnasielærernes nye kollega?

Et eksplorativt studie om brugen af AI som beslutningsstøtteværktøj til at højne korrektheden af bedømmelser af danskopgaver i gymnasiet

Semester: 6.
Periode: F24
ECTS: 15 point
Vejleder: Sarah Borup Jørgensen
Gruppenummer: 6

Antal anslag inkl. mellemrum: 186.421
Antal normalsider á 2400 tegn: 77,7
Bilag i fysiske sider: 50

Adam Alexander Faust Spies
Studienummer: 20215340

Maria Rom Kristiansen
Studienummer: 20215251

Rasmus Jakobsen
Studienummer: 20215172

AALBORG UNIVERSITET

Det Samfundsvidenskabelige Fakultet
Aalborg Universitet
Politik & Administration

Abstract

This study seeks to discover the extent to which Danish high school teachers are subjective in their grading of papers and furthermore how artificial intelligence can be helpful as a worktool in the day-to-day grading throughout the school year. The study has been conducted using an experimental design that has been implemented through a survey. The survey has been constructed through the program SurveyXact developed by Rambøll and had a total of 119 respondents who are all high school teachers in Danish on the A-level. Herein the respondents were set to grade a paper and give feedback. In addition to this a selection of the participants were given an AI support tool. The experiment was conducted through an intervention that contained feedback written on behalf of Google DeepMind's generative artificial intelligence, Gemini. Within the survey the respondents were divided into three different groups wherein two of them took part in the actual experiment and received the intervention. In addition to the quantitative method the study also performs qualitative analysis on the individual feedback that the respondents perform. Thus, the study performs mixed-method research.

To further analyze the outcome of the survey the study uses different theories from both within psychology as well as information technology. These theories are Kahneman et al.'s theory of the presence of noise in human judgement (2021) as well as the principles of the Technology Acceptance Model (TAM) as understood and perceived by Venkatesh and Davis (2000). The analysis was created around the following research question:

To what degree will the use of artificial intelligence as a decision-making-tool increase the correctness of feedback and grading within the Danish high school and specifically in the subject Danish on the A-level?

Furthermore, to specify the effects of the intervention the study created multiple hypotheses to understand and explain the outcome of the survey. These are as follows:

H1a: High school teachers that use AI as a decision-making-tool will make fewer mistakes in their grading compared to the control-group.

H1b: High school teachers that have participated as external graders on other assignments and use the AI decision-making-tool will collectively have a lower level of mistakes compared to those who have not.

H2: The group 'AI bias' will be more likely to disagree with the feedback and grading generated by artificial intelligence compared to the group 'Social Opinion Influence'.

Through analysis on behalf of the research question as well as the multiple hypothesis the study finds that the collective level of mistakes made by high school teachers are reduced when they use the AI decision-making-tool. Secondly, the teachers that have not acted as external graders become as good graders as those who have, when they use the AI tool. Thirdly, the study finds that a large group of respondents that can't side with the AI generated feedback in reality actually agree with the AI but renounce it purely due to the fact that it is artificial intelligence. This shows the importance of perceived usefulness within the Technology Acceptance Model and its crucial role to the final acceptance of a given technology. In summary the use of artificial intelligence will increase the correctness of feedback and grading within the Danish high school when looking at the subject Danish on the A-level.

Indholdsfortegnelse

1. Problemfelt

1.1 Karaktergivning

Karaktergivning er et udtryk for systematisk bedømmelse af en elev eller studerendes faglige kompetencer og færdigheder (EVA, 2016: 53). Ved hjælp af karaktergivning er potentialet for sammenligning rigtig godt, da det som udgangspunkt er den samme skala, som der bedømmes ud fra. Men hvor god en statistisk målestok er karaktergivning, hvis forudsætningerne for at opnå en given karakter ikke er ens? Det er netop denne undren, som dette projekt udspringer af.

I Danmark bliver unge præsenteret for karaktergivning i form af standpunktskarakterer allerede fra 8. klasse (UVM, 2023a). På den måde er elever bekendt med karaktergivningen, inden de potentielt skal starte på en ungdomsuddannelse.

I nyere tid er der megen debat om karaktergivning i forbindelse med det pres, der kan forekomme, når gymnasieelever skal søge ungdomsuddannelse eller videregående uddannelse med deres gennemsnit fra gymnasiet. Presset opstår som følge af, at det danske uddannelsessystem i høj grad er bygget op omkring høje gennemsnit og i mindre grad menneskelige faktorer. For at et sådant system skal kunne fungere på bedste vis, forudsætter det, at unge studerende bliver bedømt ens.

På baggrund af denne problematik vil dette projekt belyse karaktergivning for gymnasiale uddannelser.

1.1.1 7-trins-skalaen

I danske gymnasier er 7-trins-skalaen benyttet til karaktergivning siden 2006 og siden 2007 på alle øvrige uddannelser (UVM, 2023b; EVA, 2017: 5). Overgangen fra den forhenværende 13-skala skyldtes, at 7-trins-skalaen er direkte sammenlignelig med den internationale ECTS-karakterskala. Endnu engang er der fokus på det direkte sammenligningsgrundlag, som karaktergivning skaber.

Alle karakterer på 7-trins-skalaen har sin egen beskrivelse, som er gennemgående for alle fag og alle uddannelser. For at bestå, skal eleven modtage karaktererne 02, 4, 7, 10 og 12. Karaktererne og deres respektive beskrivelser ses uddybet nedenfor i tabel 1:

Karakter	Betegnelse	Beskrivelse	ECTS
12	Den fremragende præstation	Karakteren 12 gives for den fremragende præstation, der demonstrerer udtømmende opfyldelse af fagets mål, med ingen eller få uvæsentlige mangler	A
10	Den fortrinlige præstation	Karakteren 10 gives for den fortrinlige præstation, der demonstrerer omfattende opfyldelse af fagets mål, med nogle mindre væsentlige mangler	B
7	Den gode præstation	Karakteren 7 gives for den gode præstation, der demonstrerer opfyldelse af fagets mål, med en del mangler	C
4	Den jævne præstation	Karakteren 4 gives for den jævne præstation, der demonstrerer en mindre grad af opfyldelse af fagets mål, med adskillige væsentlige mangler	D
02	Den tilstrækkelige præstation	Karakteren 02 gives for den tilstrækkelige præstation, der demonstrerer den minimalt acceptable grad af opfyldelse af fagets mål	E
00	Den utilstrækkelige præstation	Karakteren 00 gives for den utilstrækkelige præstation, der ikke demonstrerer en acceptabel grad af opfyldelse af fagets mål	Fx
-3	Den ringe præstation	Karakteren -3 gives for den helt uacceptable præstation	F

Tabel 1: 7-trins-skalaen med betegnelser og beskrivelser

Det fremgår af Børne- og Undervisningsministeriets hjemmeside, at karakterskalaen skal anvendes absolut. Det vil sige, at der altid skal tages udgangspunkt i karakterbeskrivelserne set i forhold til målbeskrivelsen for uddannelsesforløbet (UVM, 2023c). Derfor udarbejdes der desuden vejledningsmateriale, hvor karakterbeskrivelser gøres fagspecifikke. Der er forskel på hvilken instans, der fastsætter 7-trins-skalaens opbygning og anvendelse på de forskellige uddannelsesinstitutioner. Dette projekt vil undersøge karaktergivning i de gymnasiale uddannelser. Her gælder det, at disse rammer er fastsat af Børne- og Undervisningsministeriet (UVM, 2020).

Sammenlignet med 13-skalaen er hensigten med 7-trins-skalaen at benytte hele skalaen i større omfang. Dette sker med et udgangspunkt i at vægte præstation med de eventuelle mangler, der kan være (UVM, 2023c). Mangler kan ikke nødvendigvis måles kvantificerbart, men er ofte af kvalitativ art, hvilket fordrer en "*betydelig faglig og fagdidaktisk indsigt, ligesom det kræver konsensus mellem bedømmerne i et givet fag i en given uddannelse, at foretage en sådan afvejning af kvalitative og/eller kvantitative mangler set i forhold til den samlede*

præstation og forløbets mål" (UVM, 2023c). Her findes det yderst relevant, at karaktergivning kræver en konsensus imellem bedømmere.
Der kan argumenteres for, at 7-trins-skalaens systematiske opbygning forekommer overskuelig. Der forekommer umiddelbart faste rammer og formuleringer for, hvornår en karakter skal gives. Dog viser det sig anderledes i praksis. Elever i overbygningen i folkeskolen peger på, at karaktergivning presser dem, og konstant inddrager dem i en intern konkurrence med sig selv og deres klassekammerater (Nielsen, 2020). Desuden angiver en karakter ikke, hvordan eleven kan udvikle og forbedre sig. Dette skal angives ved siden af i form af feedback fra bedømmeren.

I overgangen fra 13-skalaen til 7-trins-skalaen gør det sig stadig gældende, at *"En bedømmelse skal fortsat bero på et professionelt, fagligt skøn, og den skal fortsat foretages ud fra en samlet vurdering af præstationen."* (UVM, 2023c). Det professionelle, faglige skøn kan til gengæld være en åben invitation til alle de personlige præferencer, associationer, holdninger og meninger, som hver enkelt gymnasielærer måtte have om en given danskfaglig kunnen. Uden klare retningslinjer vil gymnasielærerne unægteligt foretage vurderinger, der aldrig vil være ens, fordi det i sidste ende kommer an på smag og behag (Kahneman et al., 2021: 75).

1.1.2 Bedømmelsespraksis ved karaktergivning i gymnasiet

Karaktergivning sker gennem en bedømmelsesproces. Hvordan denne proces fungerer i praksis i gymnasiet, vil blive uddybet i dette afsnit.

Dette projekts undersøgelse vil tage udgangspunkt i bedømmelsespraksissen i dansk, da karaktergivningen her ikke beror på standardiserede retteark som ved eksempelvis flere naturvidenskabelige fag. Derfor vurderes det, at karaktergivningen i dansk er ideel som genstand for denne undersøgelse.

Som en del af vejledningen for alle fag, eksisterer der læringsplaner for de danske almene gymnasier, som skitserer krav for fagets opfyldelse. Ved siden af fagets faglige mål gennemgås også de didaktiske principper såsom undervisningens progression og de ønskede receptive og produktive færdigheder. De faglige forventninger til eleven omhandler flere forskellige kompetencer. Hertil med et særligt fokus på elevens evne til at udtrykke sig præcist og have en gennemgående grammatisk forståelse. Ydermere understreges kravet om at kunne analysere, fortolke og perspektivere samtlige tekster, der præsenteres i det givne pensum. Netop dette pensum har også en række foranstaltninger. Det kræves, at gymnasieelever har en forståelse af

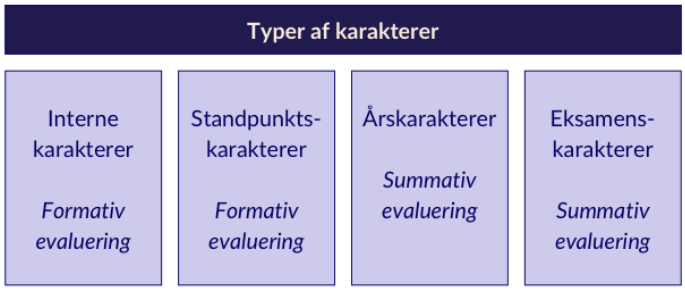
litteraturen fra flere forskellige perioder såvel som lande samt evnen til at demonstrere sit kendskab til mediebilledet og de digitale fællesskaber (UVM, 2021: 4-5). Denne opgørelse redegør for vejledning og krav til danskfaget som helhed, og det skal derfor tilføjes, at de enkelte opgaver oftest har deres egne vejledninger og krav, hvorudfra der gives en karakter. Denne opgørelse af faget skal ses som den samlede forventning af, hvad en elev i dansk på A-niveau skal kunne præstere for at opfylde kravene til faget.

Ved bedømmelsespraksissen skelnes der mellem to typer af evaluering. Disse er henholdsvis formativ- og summativ evaluering. Den formative evaluering dækker over den løbende vurdering af elevens præstation holdt op imod delmål med henblik på at understøtte den videre læringsproces. Det vil eksempelvis sige standpunktskarakterer og karaktergivning i forbindelse med enkelte opgaver og præstationer. Den summative vurdering rummer karaktergivning i forbindelse med en endelig præstation (EVA, 2016). Uanset hvilken form for evaluering og karaktergivning, der er tale om, er det afgørende for troværdigheden af karaktergivning, at dette sker ud fra de samme forudsætninger.

I løbet af gymnasietiden modtager elever forskellige former for karakterer, som rummer begge typer evaluering. Den hyppigste af disse er den interne karakter, der gives ved de lokale opgaver og afleveringer igennem uddannelsen. Her testes eleverne i dele af den samlede læringsplan for bedst muligt, at kunne opfylde alle krav til faget, når den endelige prøve skal tages. En opgave kan eksempelvis tage udgangspunkt i en given skrivegenre eller en tidsperiode, der er relevant for den samlede læringsplan i dansk på A-niveau. Derved fungerer de interne karakterer som en slags pejlemærker for eleven, hvorfor denne type karakter er et udtryk for formativ evaluering.

Derudover er der standpunktskarakteren, der skal give eleven en vurdering af dennes foreløbige faglige standpunkt. Denne karakter gives løbende i skoleåret. Den gives i de fag, som eleven ikke færdiggør, men fortsat skal undervises i året efter, hvorfor denne også er et udtryk for formativ evaluering (EVA, 2016: 27). Modsat standpunktskarakteren er årskarakteren den endelige vurdering af elevens faglige niveau ved afslutningen af skoleåret, hvorfor denne både er et udtryk for summativ- og formativ evaluering. Som den sidste evaluering af fagets opfyldelse er det også denne karakter, som ender på karakterbladet for eleven. Slutteligt er prøvekarakteren, der forekommer i sideløb med den afsluttende årskarakter. Denne karakter gives til de afsluttende prøver/eksaminer, når et fag færdiggøres. Denne karakter vil også forekomme på karakterbeviset sammen med årskarakteren og er også et udtryk for summativ evaluering. Det er de karakterer, som udgøres af den summative

evaluering, der benyttes, når elever skal søge enten ungdomsuddannelse eller en videregående uddannelse. Derfor kan disse siges at bruges som en slags selektionsredskab for uddannelsesinstitutioner i forbindelse med optag på samme (UVM, 2020).
De fornævnte karaktertyper ses visualiseret i nedenstående tabel 2:



Tabel 2: Typer af karakterer

Ved eksamener bedømmes skriftlige opgaver ved ekstern censur. Denne eksterne censur består som udgangspunkt af ansatte ved gymnasiale institutioner, men kan i særlige tilfælde også være beskikkede eksterne censorer (UVM, 2024a).

Gymnasielærere byder ved skoleårets begyndelse ind på, hvor mange opgaver, de ønsker at rette ved skriftlige prøver. Forud er der opgivet normer for skriftlig censur ved de gymnasiale uddannelser, som danner ramme for, hvor mange opgaver det forventes, at man som censor kan nå at rette på en time. For dansk A på stx, hhx, htx og hf-2 er det eksempelvis opgivet, at normen er to rettede opgaver per time (UVM, 2023d). Censorer har desuden mulighed for at blive vejledt af fagkonsulenten og deltage i kurser, der har til formål at klæde censoren på til opgaven. Dette skal sikre, at alle censorer retter opgaverne på samme måde, og at bedømmelserne bliver så nøjagtige og korrekte som muligt (Styrelsen for Undervisning og Kvalitet, 2024: 1).

Hver eksamensbesvarelse rettes af to eksterne censorer, som tildeles fra centralt hold. Det er ikke muligt at få sin egen underviser eller andre undervisere fra ens institution som censor. Alle censorer mødes på censormødet, hvor besvarelserne voteres og der angives karakter (UVM, 2024b).

Anderledes er det ved de formative evalueringer, som foretages i løbet af året af elevens egen lærer. Her er det som udgangspunkt udelukkende elevens egen lærer, som alene foretager

vurderinger af elevens præstationer. Der skelnes desuden ikke mellem censorer og dem, der ikke er det.

Dertil er de formative evalueringer anderledes fra de summative, da læreren kender hver opgaves skribent. På den måde kan læreren tage udgangspunkt i elevens proces og udvikling, og foretager sin vurdering ud fra dette. Dog fordrer dette, at læreren vurderer sine elever på det samme grundlag. Dertil fremhæves igen, at der skal være en konsensus imellem bedømmere i forhold til vægtning af fejl og styrker ved karaktergivning (UVM, 2023c). Er der ikke det, så forsvinder idéen med et kvantitativt karaktersystem.

1.1.3 Kvalitet i karaktergivning

Nutidens system for videregående uddannelser er i stor udstrækning bygget op omkring adgangskvotienter, som udgøres af et vægtet gennemsnit af studerendes eksamensbevis fra gymnasiet. Derfor er det væsentligt, at standarden for bedømmelsespraksissen er ens i hele landet. Det kan potentielt betyde en drømmeuddannelse til forskel, hvorvidt en elev er bedømt gavmildt eller strengt.

En analyse fra CEPOS (2021), skildrer en problematisk og gennemgående forskel mellem gymnasieelevers karakterer gennem året samt til eksamen. Af analysen fremgår det, at både køn, etnisk herkomst og geografi spiller en beviselig rolle i forbindelse med karaktergivning (Kjeldsen & Larsen, 2021: 1-3). Generelt forekommer det, at alle grupper får højere årskarakterer end prøvekarakterer. Det er problematisk, hvis gymnasieelever generelt ikke kan gå ud fra, at de i en eksamenssituation ligger på samme niveau som gennem året. Desuden vidner dette om, at ikke alle gymnasielærere lægger vægt på de samme kriterier, når de retter opgaver.

Endvidere spekuleres der i analysen i, hvorvidt der forskelsbehandles i forbindelse med karaktergivning i årets løb. Eksempelvis får ikke-etnisk danskere i gennemsnit en prøvekarakter, der er 1,08 karakterpoint højere end den, de får til eksamen (Kjeldsen & Larsen, 2021: 1). Analysen påpeger yderligere, hvorvidt en anderledes udformning af bedømmelsessystemet vil kunne ændre denne forskel på tværs af grupper.

Problematikken angående den manglende konsensus i karaktergivning blev yderligere skildret i en bachelorafhandling fra 2022, hvor den teoretiske ramme baserede sig på Kahneman et al.s (2021) teori om støj i bedømmelser. Denne viste, at den samme danskopgave i folkeskolen

kunne få alle karakterer fra 02 til 12 fra undersøgelsens 51 forskellige bedømmere (Roersen et al., 2022).
Dette skildrer et paradoks, hvor karakterer på den ene side er udslagsgivende for, hvilke uddannelser unge studerende kan søge ind på, og hvilke de ikke har mulighed for at søge, samtidig med, at der forekommer subjektive bedømmelser, som skaber en vis tilfældighed i karaktergivningen.

Dette lader dog ikke til at være den eneste udfordring i gymnasierne. Mellem 2016 og 2019 oplevede gymnasierne besparelser på 1,5 milliarder kroner. Disse besparelser har ifølge Gymnasieskolernes Lærerforening resulteret i en forringelse af undervisningskvaliteten for eleverne (Gymnasieskolernes Lærerforening, 2019: 1). Der er mindre tid til at differentiere undervisningen, mindre tid til den enkelte elev og færre ressourcer til de dygtige elever. Kun en fjerdedel af lærerne kan stå inde for den undervisning, som de eksisterende rammer muliggør (Gymnasieskolernes Lærerforening, 2019: 4). Det berettes, at eleverne efterspørger mere konkret feedback på skriftlige opgaver, men grundet det stigende antal skriftlige opgaver og den mindre tid er kvaliteten faldet på både feedbacken og lærernes forberedelse til undervisningen (Gymnasieskolernes Lærerforening, 2019: 7). Alt i alt har besparelserne på området ført til mindre feedback på skriftlige afleveringer, mindre gennemarbejdet undervisning og mindre fokus på elevernes individuelle behov (Gymnasieskolernes Lærerforening, 2019: 11).

1.2 National strategi for digitalisering

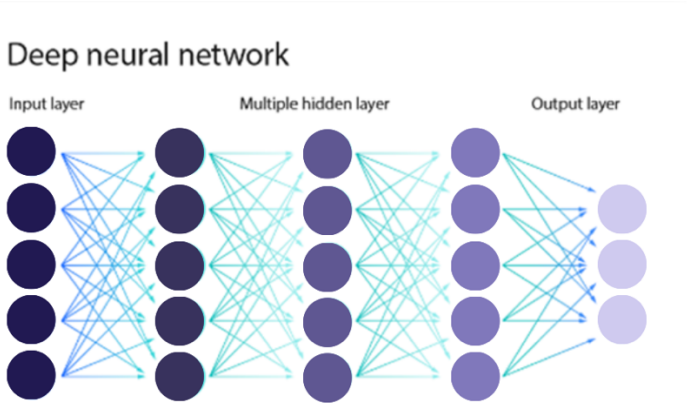
Med viden om, hvilke problemstillinger og udfordringer, der er angående kvaliteten af karaktergivning og feedback af opgaver i gymnasiet, er det nærtliggende at undersøge, hvordan disse problemer kan løses. En mulighed kunne være at udnytte de store teknologiske fremskridt, som sker i samfundet omkring os.
I disse år går den teknologiske udvikling stærkt. Især inden for de seneste årtier har der været fokus på digitaliseringen af samfundet. Det er for borgere blevet hverdag at kunne tilgå virksomheder og offentlige myndigheder på internettet. Også på arbejdsmarkedet har teknologien ændret arbejdsgange, lettet arbejdsbyrder og effektiviseret processer, men udviklingen fortsætter. Her taler forskere om 'Den fjerde industrielle revolution', hvilket indbefatter, at kunstig intelligens bliver den nyeste medarbejder og løser arbejdsopgaver, som mennesker før varetog (Krause-Jensen et al., 2022: 5).

Regeringen har siden den fællesoffentlige digitaliseringsstrategi fra 2001 haft en national strategi for digital udvikling. Senest i maj 2022 udkom SVM-regeringens strategi for digitalisering i både den offentlige og private sektor. Heri er en række punkter, hvor regeringen har som ambition at udnytte teknologien bedre for at sikre Danmarks digitale fremtid (Finansministeriet, 2022: 3). Regeringen vil i den forbindelse investere to milliarder kroner i digital udvikling over de næste fem år. Investeringerne skal være med til blandt andet at omstille velfærdssamfundet, udnytte digitale redskaber bedre, sikre os mod trusler udefra og accelerere den grønne omstilling (Finansministeriet, 2022: 3). Regeringen skriver endvidere, at digitaliseringen skal skabe værdi for borgere ved at udvikle samfundet og løse konkrete problemstillinger og udfordringer (Finansministeriet, 2022: 12).
De konkrete problemstillinger og udfordringer i gymnasiet kan potentielt løses gennem bedre udnyttelse af teknologi og digitalisering, herunder anvendelse af kunstig intelligens.

1.3 Kunstig intelligens

Kunstig intelligens er hurtigt blevet et anvendt begreb af mange i deres daglige tale. Årsagen til dette er blandt andet lanceringen af chatbotten ChatGPT, der siden dens lancering i november 2022, har været underlagt megen debat. ChatGPT blev sågar også kåret til årets ord i 2023 (Dansk Sprognævn, 2023). Chatbots som denne er dog ikke lig med kunstig intelligens, men er blot en afgrening af det. De anvender såkaldt *'deep learning'* som giver systemet evnen til at lære data uden programmering. Ud fra dette forsøger chatbotten at imitere processen fra en menneskelig hjerne. Den gennemgår tre forskellige faser eller lag i dens proces: et input lag, et skjult lag (laget hvor robottens egen viden er) og et output lag. Den anvender algoritmer alt efter det input den har modtaget, hvorefter den identificerer sin data og bruger dens fund til at generere et output (IBM, 2024).

Processen for *'deep learning'* chatbots er skitseret i nedenstående figur 1 (IBM, 2024).



Figur 1: Det neurale netværk for deep learning chatbots

Med udgangspunkt i dette netværk er den nye form for generativ kunstig intelligens blevet til. Denne sprogmodel eller GPT (Generative Pre-Trained Transformer), som er udviklet af OpenAI, har evnen til at genere tekst og udføre sproglige opgaver. Den er trænet i at anvende forskellige data fra diverse hjemmesider for at give det bedst mulige svar-output til det input eller prompt, den er givet. Den nuværende udgave af GPT (GPT-3) anvender 175 milliarder læringsparametre til at skrive sit svar (Floridi & Chiriatti, 2020). Disse parametre skal forstås som robottens egen viden og bruges til at generere nye data, der ligner det lærte, men ikke er identisk med det (IBM, 2024).

Netop denne type af kunstig intelligens har været årsag til store debatter omkring anvendelse i det danske skolesystem. Den er især i rampelyset, når der tales om snyd og plagiat i forbindelse med, at eleverne skal skrive opgaver. Det er samtidig en debat, der deler lærerne. En gymnasielærer udtaler sig: *"Vi skal være modige og nysgerrige i forhold til at inddrage kunstig intelligens, for vi kan ikke forbyde os ud af det"*, i den anden grøft siger en gymnasielærer: *"Jeg synes ikke, at ChatGPT er en særlig kompetent samtalepartner – og slet ikke fagligt. Den er alt for biased"*. Og denne sidste beretning er en som mange danske gymnasielærere deler. Der er stor frygt for læringsmæssige tab, når eleverne anvender generativ kunstig intelligens i skolen. Trods det ikke er alle gymnasielærere, der er enige i modstanden af generativ kunstig intelligens, er det klart den største gruppering i debatten. (Romme-Mølby, 2023).

Det store problem ved denne type af generativ kunstig intelligens er, at det grundet de enorme mængder data bag, ikke altid er muligt at identificere processen for, hvordan den er kommet frem til sit output. Det er altså ikke muligt at begrunde det svar, der er givet, hvor der opstår problematikken om den sorte boks. Her skal den sorte boks forstås som det skjulte lag i det neurale netværk. Hvis chatbotten eksempelvis genererer et output om sundhed vil det ikke være muligt at garantere, hvorvidt det er baseret på fagproffessionel viden eller subjektive holdninger (IBM, 2024).

1.4 Beslutningsstøtteværktøj

Hvis man forsøger at løse problemerne ved feedback og karaktergivning i gymnasiet ved brugen af kunstig intelligens, så kunne det minde om det, man i fagsprog kalder for et beslutningsstøtteværktøj. Beslutningsstøtteværktøj er en fællesbetegnelse for redskaber, der kan forberede og hjælpe fagprofessionelle til at foretage den bedste beslutning i givne situationer. Ofte er beslutningsstøtteværktøjer nævnt i forbindelse med klinisk beslutningsstøtte. Eksempelvis opererer hospitaler i Danmark ofte med det generiske beslutningsstøtteværktøj *Ottawa Personal Decision Guide*, der hjælper sundhedsprofessionelle med at bestemme behandlingsvalg i samarbejde med patienten. Dette sikrer, at patienten har en visuel oversigt over hvilken behandlingsmulighed, der har flest af de fordele, som er vigtige for ham eller hende og færrest af de ulemper, som patienten bekymrer sig mest om (Danske Patienter, 2024). Nogle beslutningsstøtteværktøjer er baseret på algoritmer eller retningslinjer, mens andre er formidlingsredskaber til brug i dialogen med patienten om netop præferencesensitive beslutninger (Lindebjerg & Rahr, 2017).

I tilfældet, hvor gymnasielærere skal bedømme danskopgaver, vil der uden tvivl være tale om beslutninger, der baserer sig på præferencer. Her vil kunstig intelligens muligvis kunne støtte gymnasielærerens fagprofessionelle skøn og sikre, at beslutningen om karakteren er så korrekt som overhovedet muligt.

1.5 Problemformulering

Gennem problemfeltet er der skitseret et paradoks bestående i, at karaktergivning i gymnasiet er meget afgørende for elevers fremtidsmuligheder, men at der forekommer subjektivitet i karaktergivning på tværs af køn, region og etnicitet (Kjeldsen & Larsen, 2021). Desuden har

nedskæringer på gymnasierne resulteret i lavere kvalitet af feedback og mindre tid til at rette skriftlige opgaver - som der i øvrigt bliver flere og flere af. Dette skaber et større rum for fejlagtige karakterer, når det gælder den faglige vurdering af den enkelte elev. Desuden er der nye muligheder i form af generativ kunstig intelligens, der kan hjælpe digitaliseringen af skolesystemet på vej og herunder anvendes ved karaktergivning på de danske gymnasier.

På baggrund af problemfeltets redegørelser og med de ovenstående observationer in mente udleder projektet følgende problemformulering:

I hvilken udstrækning vil brugen af AI som beslutningsstøtteværktøj højne korrektheden i feedback- og karaktergivning i skriftlig dansk på A-niveau i gymnasiet?

1.6 Afgrænsning

Projektets problemformulering afgrænses hermed til et spørgsmål, der har til formål at besvare, men samtidig undersøge, hvorvidt et beslutningsstøtteværktøj bygget på kunstig intelligens kan mindske fejl i gymnasielærernes karaktergivning. Når det gælder debatten om karaktergivning i danske gymnasier, er der forskellige perspektiver, som for udarbejdelsen af denne undersøgelse, må afgrænses. Problemstillingen er meget kontekstafhængig, da den er baseret på forventningen om, at det mønster i karaktergivning, som rapporten fra CEPOS (2021), samt tilstedeværelsen af støj ved karaktergivningen som afhandlingen af Roersen et al. (2022) påpeger, også er gældende for dette projekts undersøgelse så vel som respondenterne. Da denne undersøgelse vil foregå ved brug af mixed-method, vil det både være muligt at skitsere de overordnede indvirkninger fra den udvalgte teori såvel som årsagerne bag lærernes individuelle bedømmelser. Der er valgt at afgrænse projektet således, at der tages udgangspunkt i gymnasielærere, der underviser i dansk på A-niveau. Dette valg tages først og fremmest, da det vil være muligt at anvende alle de sproglige værktøjer, som en moderne sprogmodel har. Ydermere er dansk A også et fag som er udbredt på de danske gymnasier. Derved vil projektets undersøgelse kunne omfavne flest mulige lærere og dermed implicit også deres elever. Samtidig vil undersøgelsens målingsvaliditet også højnes af dette valg.

Det output, der skabes ved hjælp af kunstig intelligens, vil heller ikke kunne blive perfekt. Som nævnt i problemfeltet vil dette projekt være udfordret i form af den sorte boks. Det vil forsøges at prompte den udvalgte sprogmodel således, at den tager udgangspunkt i den nedskrevne faglighed for dansk på A-niveau. Dog vil det ikke være muligt at fjerne den

eksisterende data i modellens usynlige lag. Derfor må det endelige output anses som en prototype i undersøgelsens forsøg.

Slutteligt vil det for projektets undersøgelse ikke være muligt at redegøre for, hvor 'forkert' en karakter er. På baggrund af teorien om fejl i bedømmelser og den dertilhørende metodes udformning, vil der blive sat en sand værdi, der er den "korrekte" karakter, som er genereret af det kunstige output. Dette er også gældende for de karakterer, som den udvalgte opgave bliver tildelt af respondenterne. Afstanden fra karaktererne 00 til 7 er på papiret, næsten den samme som fra 7 til 12. Men et 10- eller 12-tal har en helt anden vægtning end et 00-, 02- eller 4-tal har. Derfor vil projektet i sin analyse af undersøgelsen se bort fra denne forskel og måle resultaternes variation i forhold til den sande værdi som kunstig intelligens producerer.

2. Litteraturreview

I dette afsnit vil den eksisterende forskning inden for undersøgelsens genstandsfelt præsenteres. Samtidig vil det også kortlægges, hvorledes undersøgelsen afviger fra den eksisterende forskning og bidrager til forskningsfeltet. Ydermere vil den anvendte litteratur, der har skabt grundlaget for undersøgelsen, præsenteres og belyse udgangspunktet for analysen og senere sættes i relief med den udvalgte teori. Litteraturgennemgangen vil derved skabe afsættet for denne bachelorafhandlings empiriske substans og analytiske grundlag.

2.1 Gennemgang af litteratur

Det anvendte bachelorprojekt er fundet i Aalborg Universitets projektbibliotek. De anvendte artikler og andet empiri er fundet gennem Aalborg Universitetsbibliotek, Google Scholar og almen litteratursøgning. Litteraturen er udvalgt på baggrund af dens relevans for undersøgelsens videnshul og som hjælp til at belyse og skitsere de problematikker, analysen vil undersøge. Først vil problematikken omhandlende karaktergivning i danske gymnasier samt arbejdspresset på gymnasielærerne selv skitseres. Herefter vil den eksisterende forskning præsenteres, den udvalgte litteraturs anvendelse vil skitseres, og det vil blive beskrevet hvordan denne bachelorafhandling bidrager til feltet.

2.2 Eksisterende forskning

Karaktergivningen i de danske gymnasier har siden indførelsen af 7-trins-skalaen været omdiskuteret fra mange forskellige vinkler. Som grundlag for undersøgelsens problemformulering anvendes rapporter fra henholdsvis Danmark Evalueringsinstitut (EVA) og tænketanken CEPOS. Disse rapporter undersøger diverse problematikker omhandlende karaktergivning i de danske gymnasier.

I 2020 udgav EVA en rapport, der opgjorde resultaterne af et forsøg med forskellige gymnasier, der undlod at give karakter til deres respektive 1.g-klasser. Forsøget blev foretaget på baggrund af implementeringen af gymnasiereformen i 2016 og blev udført på årgangen, der startede i 2017. Ønsket med rapporten og forsøget var at skabe viden om, hvordan man højnede feedbacken og hele karaktersystemet (EVA, 2020). Forsøget fandt, at to ud af tre elever i forsøget med et karakterfri 1.g var overvejende tilfredse med ordningen. Tilfredsheden

varierede dog på tværs af skoler og ud fra elevernes baggrund. Eleverne ønsker, ifølge rapporten, hyppig feedback, men også karakterer en gang imellem. De har et ambivalent forhold til karakterer, da de både oplever usikkerhed uden dem, men også ser dem som en anerkendelse og en del af deres identitet. Forsøget har vist, at der er behov for at finde en balance mellem karakterer og feedback, der tager højde for elevernes forskellige behov. Ydermere foretog de interviews med 15 lærere. Herfra kunne det udledes, at lærerne i deres egen vurdering føler, at de står meget alene med opgaven om at oversætte de faglige mål til en helhedsvurdering i form af karakteren. Årsagen er, at de faglige mål ofte er beskrevet overordnet og er for abstrakte. Dette leder derfor op til personlig og subjektiv fortolkning af hvilke kompetencer, der kræves til den enkelte karakter. Lærerne beretter også, at de oftest drøfter karaktergivning uformelt og ad hoc eksempelvis over frokosten (EVA, 2020).

Året efter, i 2021, udgav CEPOS en analyse, der omhandlede forskelle i karakterer på de gymnasiale uddannelser. Analysen viste, at der var store forskelle mellem de skriftlige års- og prøvekarakterer på danske gymnasier. Forskellene varierer især mellem mænd og kvinder, samt mellem danske elever og elever af anden etnisk herkomst end dansk. CEPOS fandt denne store forskelsbehandling og vurdering yderst bekymrende, da dette udfordrer retfærdigheden og objektiviteten i det danske karaktersystem. Ydermere fandt analysen flere problematikker omhandlende udformningen af bedømmelsessystemet (Kjeldsen & Larsen, 2021).

I et studie af, hvordan undervisere forholdte sig til den store stigning i online undervisning under COVID-19 anvender Hong, Zhang og Liu Teknologi Accept Modellen (TAM) til at undersøge de afgørende årsager til den endelige anvendelse. Deres undersøgelse blev foretaget på baggrund af surveydata fra 1.568 undervisere under COVID-19 pandemien. Undersøgelsen fandt en tydelig sammenhæng mellem teknologiens brugervenlighed, brugbarhed såvel som jobrelevans og den endelige brugeradfærd angående online undervisning. De fandt yderligere derigennem tydelig anvendelsesmulighed for TAM til undersøgelse af kinesiske undervisere, der står i nødsituationer. Slutteligt fremhæver deres resultater mulighederne for interventioner, der har til formål at forbedre underviseres accept af teknologi til brug i undervisningen (Hong, Zhang og Liu, 2021).

Et andet studie af Cabitz et al. (2023) undersøgte, hvordan AI ville dominere beslutningsprocessen inden for fagprofessionelle sundhedsvurderinger i forskellige situationer som eksempelvis aflæsning af et røntgenfotografi eller lignende medicinske prøver. I deres eksperiment af, hvorvidt AI ville have en negativ indvirkning på den endelige bedømmelse af

de udvalgte cases, fandt de en statistisk stigning i, hvad de kalder 'gavnlig selvtillid' (*beneficial self-reliance*) og et fald i 'skadelig overtillid' (*detrimental over-reliance*). Dette understøttede deres studie yderligere, da det fjernede det såkaldte *'automation bias'*, altså skævheder ved hyppig anvendelse af deres AI løsning. De individer som deltog i deres undersøgelse faldt ofte tilbage på deres egne fagprofessionelle skøn i de tilfælde, hvor AI foregav fejlagtig støtte, og det var derfor muligt for dem, at fraskrive hypotesen om AI dominans ved anvendelse af deres støtteværktøj (Cabitz et al., 2023).

Slutteligt drager denne undersøgelse også inspiration fra en anden bachelorafhandling fra 2022. Studiet undersøger betydningen af bedømmelsespraksissen i dansk i folkeskolens for karakterernes troværdighed. Studiet fandt, at karaktererne er vigtige for unge, men der er udfordringer i form af karakterernes ensartethed. Studiet bad 51 lærere om at bedømme en danskopgave fra en 9. klasses prøve. De blev inddelt i to grupper, hvor den ene gruppe bedømte som de plejer, og den anden gruppe bedømte efter seks kriterier. Dette forsøg viste, at den analytiske bedømmelse, der foregik på baggrund af de seks kriterier, markant reducerede graden af støj i lærernes karaktergivning, jævnfør teori af Kahneman et al. (2021). Slutteligt viste studiet, at de lærere der havde erfaring som censor, eller havde været ansat som lærer i mere end 10 år, havde samme lave påvirkning (Roersen, et al., 2022).

2.3 Anvendelse af litteraturen

Med den eksisterende litteratur in mente og med afsæt i den udvalgte forskning, kan der opstilles en tydelig problematik omhandlende bedømmelsespraksissen for danske gymnasielærere såvel som de muligheder, AI giver for understøttelse af fagprofessionaliteten. Årsagen til problematikken synes af projektgruppen at rumme flere forskellige begrundelser. Der forekommer et tydeligt billede af, at den tid, der forventes at være til rådighed for den enkelte gymnasielærer, ikke er anvendelig i praksis. Ydermere tegner der sig et billede af, at gymnasielærer mangler konsensus i bedømmelserne. Derudover er der en ulighed mellem forventningerne og de besparelser, som gymnasierne har oplevet som resultat af gymnasiereformen. Slutteligt giver den eksisterende forskning et billede af mulighederne som kunstig intelligens skaber inden for forskellige erhverv og dertil også, at det muligvis ikke er så farligt eller upålideligt i samarbejde med fagprofessionaliteten, som de fleste tror.

2.4 Afhandlingens bidrag

Der er lavet meget forskning om både de danske gymnasier samt bedømmelsespraksis på tværs af hele uddannelsessektoren. Men med introduktionen af store sprogmodeller og *'deep learning'* chatbot systemer, præsenteres nye og anderledes muligheder for, hvordan det danske undervisningssystem kan fungere i praksis. Der er heller ikke eksisterende forskning eller dybdegående undersøgelser, som tager udgangspunkt i den rivende udvikling, der sker i digitalisering lige nu. Samtidig gør denne afhandlings forskningsdesign undersøgelsen unik, da den ved hjælp af et større surveyeksperiment både kan kontrollere, hvorvidt AI som beslutningstøtteværktøj højner korrektheden af karaktergivning, mens det desuden er muligt at klarlægge hvordan de danske gymnasielærere forholder og tilslutter sig til kunstig intelligens som et beslutningsstøtteværktøj. Ved at opdele lærerne i tre forskellige grupper vil det kunne skitseres, hvorvidt bias, støj og/eller teknologi benægtelse har en indvirkning på gymnasielæreres holdning til implementeringen af kunstig intelligens som værktøj. Ved at operationalisere den udvalgte teori på de indsamlede besvarelser, vil det videnshul, som projektgruppen har fundet, kunne afdækkes.

Afhandlingen vil derfor kunne bidrage til forskningsfeltet ved først og fremmest at belyse praksissen i gymnasielærernes proces, når de vurderer opgaver og giver karakter. Dernæst vil den, med udgangspunkt i analysen, kunne skitsere hvilken indvirkning Kahneman et al.s (2021) teori om støj har på det endelige teknologiske produkt. Der skabes yderligere en dybere forståelse af hvilke faktorer i Teknologi Accept Modellen (Venkatesh & Davis, 2000), der har en afgørende betydning for, hvorvidt danske gymnasielærere vil kunne se sig selv anvende en udlært sprogmodel som værktøj i deres bedømmelsespraksis. Kulminationen af alle disse fund vil slutteligt kunne udforme sig i et bud på, hvorvidt en bedre udnyttelse af teknologi på de danske gymnasiale uddannelser vil højne korrektheden af bedømmelser.

3. Teori

3.1 Fejl i bedømmelser

Når en gymnasielærer læser en danskopgave og efterfølgende skal give opgaven feedback og en passende karakter, går en del mekanismer i gang i løbet af processen forud for karakterafgivelsen. Beslutningen om hvilken karakter opgaven fortjener, kan være meget vigtig. Det viser sig dog, at beslutninger, hvor professionelle vurderer og bedømmer på baggrund af dømmekraft, minder om et lotteri (Kahneman et al., 2021: 53). I det følgende vil det blive gennemgået, hvad en bedømmelse er, hvorfor processen ofte ikke er optimal, hvad der foregår under bedømmelsen og slutteligt, hvordan man kan gøre bedømmelser mere objektive.

3.1.1 Vurderende dømmekraft

En beslutning kan være mange ting; hvilket job du skal søge, hvilket hus du vil byde på eller hvem du vil giftes med? I samfundet omkring os tages der også beslutninger; hvilken dom skal den anklagede have, hvem skal vi ansætte i firmaet eller hvilken score skal gymnasten have (Kahneman et al., 2021: 35; 51)? I den perfekte verden er alle beslutninger retvisende (Kahneman et al., 2021: 3). Dommen til den anklagede skal matche alvorligheden af forbrydelsen lige så vel som elevens karakter skal matche niveauet i opgaven. Men i alle beslutninger er der dog stor risiko for, at fejl sker i forbindelse med beslutningsprocessen hvilket betyder, at ikke alle bedømmelser rammer plet (Kahneman et al., 2021: 3).

Når der er tale om vurderinger, hvor der ikke er et endeligt mål at sigte efter, så må det erkendes, at der i alle vurderinger er en forventning om uenighed. Selv hvis alle bedømmere får adgang til præcis de samme oplysninger, så vil bedømmerne aldrig være helt enige om bedømmelsen, fordi bedømmelser beror på præferencer og værdier. Én ting er dog sikkert; bedømmerne vil alle hver især være selvsikre om, at netop deres bedømmelse er den rigtige (Kahneman et al., 2021: 52). Men hvis to lærere er uenige om, hvilken karakter eleven skal have, så må mindst én af dem tage fejl. Fejlen kan skyldes mange ting. Måske er den ene lærere mindre dygtig end den anden. Måske er læreren tilmed udsat for forskellige former for støj i forbindelse med bedømmelsesprocessen. Andre eksempler er mere ekstreme: Den anklagede, hvis forbrydelse bliver takseret til forskellige grader af domme, er udsat for det som Kahneman et al. kalder for vilkårlige grusomheder. At elevens opgave kan blive bedømt forskelligt afhængigt af hvilken censor, der bedømmer den og dermed potentielt kan slukke drømmen om

at søge ind på drømmestudiet, er forkert. Uretfærdighed forårsaget af systemstøj er forkert. Det skader troværdigheden til systemet (Kahneman et al., 2021: 53). systemstøj vil blive redegjort for senere.

3.1.2 Bias

Der foregår forskellige psykologiske processer i hjernen i forbindelse med bedømmelser. Ifølge Kahneman et al. (2021), er heuristikker og bias en stor del af individers beslutningsproces. Heuristikker er kognitive genveje, der hjælper individer til at simplificere svære spørgsmål, så beslutningsprocessen lettes (Kahneman et al., 2021: 161). I den forbindelse taler Kahneman et al. (2021) om, at hjernens tænkning er inddelt i to adskilte systemer. System-1-tænkningen finder hurtigt og intuitivt løsningsforslag til problemer, der er baseret på associationer (Kahneman et al., 2021: 182). Disse intuitive løsninger leverer ofte tilstrækkeligt med udbytte til de fleste svar, men de kan dog føre til bias i beslutningsprocessen (Kahneman et al., 2021: 161). System-2-tænkningen er herefter godkendelsen af de første intuitive tanker, der er opstået. Denne proces er langsommere og mere reflekterende. System-1-tænkningen skal godkendes, før den kan kaldes for en decideret holdning til en given sag (Kahneman et al., 2021: 182). I mellemtiden kan der dog opstå bias, hvilket kan føre til fejl i den endelige beslutning eller bedømmelse.

Bias er defineret af Kahneman et al. (2021) som "*psykologiske mekanismer, der fører til fejl i bedømmelser*". Disse fejl er specifikke og identificerbare (Kahneman et al., 2021: 163-164). Oftest sker bias i beslutningsprocesser i én retning. Eksempelvis er biaset kaldet 'Planlægningsfejlslutningen', når en person skal give et estimat over, hvornår man er færdig med en given opgave, så vil tidsestimatet oftest være for optimistisk end det reelle tidsforbrug, der er nødvendigt. Dette betyder, at når den sande værdi af en bedømmelse er ukendt, så vil den endelige bedømmelse være præget af psykologisk bias (Kahneman et al., 2021: 162).

Der findes mange forskellige typer af bias, som kan føre til fejl i beslutningsprocesser for individer. Følgende vil være en beskrivelse af de bias, som er relevante for projektet:

konklusionsbias starter i system-1-tænkningen. Når individer bliver præsenteret for et problem, en opgave, et udsagn eller lignende, så vil individet altid have en fordom. Som nævnt tidligere arbejder system-1-tænkningen med associationer, der giver hurtige løsningsforslag. Fordommen, som er indlejret i system-1-tænkningen, bevirker at starten af bedømmelsesprocessen allerede hælder til en bestemt konklusion. Herefter udspilles ét af to

mulige scenarier: 1) Individet springer direkte til konklusionen, eller 2) system-2-tænkningen mobiliseres således, at den finder argumenter, der matcher system-1-tænkningen og fordommen fra starten. Dette sker ved at hjernen indsamler den information, som passer til den fordom, som konklusionen beror på, og som individet håber er rigtig (Kahneman et al., 2021: 169). Ofte er styrken af denne bias forbundet med de følelser, som individet har om sagen.

Et tænkt eksempel er, at et individ har en fordom om, at kunstig intelligens er skadeligt for vores samfund og menneskets eksistensgrundlag. Individet bliver nærmest irriteret, når det hører om, at AI vinder indpas flere og flere steder i samfundet. Når individet bliver præsenteret for evidens, der bakker om om kunstig intelligens og dets potentiale til at forandre verden i en positiv retning, så vil dets følelser om kunstig intelligens og fordomme om samme, højest sandsynligt påvirke individets bedømmelse af dette udsagn. Det vil højest sandsynligt få individet til at springe til konklusionen om, at det er uenigt og ikke tror på, at udsagnet er sandt. Dermed har det allerede en forvrænget bias i en bestemt retning (Kahneman et al., 2021: 169-170).

Førstehåndsindtryk og fordomme former sig hurtigt og er meget svære at ændre igen. Ifølge Kahneman et al. (2021) skaber individer overdreven sammenhæng mellem deres fordomme og deres konklusioner. Denne overdrevne sammenhæng får individer til at tilsidesætte al modsatrettet evidens for at holde fast i fordom og førstehåndsindtryk (Kahneman et al., 2021: 172). Helt naturligt regner individer med, at den viden, som holdningen bygger på, er valid, men den er ofte forvrænget til at passe til system-1-tænkningen. Her refereres endnu engang til associationer og information, som individet har hørt eller set. Disse informationer er svære at glemme og umulige at ignorere i en beslutningsproces (Kahneman et al., 2021: 173).

Både konklusionsbias og overdreven sammenhæng skaber systematisk bias, som er konsistent skævhed i bedømmelser (Kahneman et al., 2021: 4). Systematisk bias kan måles. Støj, som er usynlig og tilfældig er sværere at måle. Det er variationen i bias mellem mennesker, der skaber støj. Individuelle forskelle i bias fra sag til sag skaber ydermere bias, der fører til flere overdrevne sammenhænge, som fører til endnu flere forvrængede slutbedømmelser. Dette kan kædes sammen med systemstøj (Kahneman et al., 2021: 174). Derfor mener Kahneman et al. (2021), at alle foranstaltninger, der reducerer bias, vil føre til bedre og mere objektive bedømmelser.

3.1.3 Støj

Hvor bias fører til systematiske fejl, der er specifikke og identificerbare, så er fejl forårsaget af støj fuldstændigt tilfældige. Her findes ingen forklaring på afvigelserne fra målet (Kahneman et al., 2021: 4). Der er dog ifølge Kahneman et al. (2021) flere forskellige typer af støj, der påvirker individers dømmekraft.

3.1.4 Niveaustøj

Når nævninge i en retssag skal beslutte hvilken dom, de synes, er passende til den anklagedes begåede forbrydelse, så er det usandsynligt, at de er enige alle sammen. Selvom de har adgang til den samme mængde information, straffer nogle nævninge hårdt, mens andre nævninge er mildere. Dette kaldes niveaustøj. Disse forskelle i bedømmelserne afspejler variationen imellem nævningenes personlige karakteristika og har intet med retfærdighed at gøre. Disse kunne være nævningenes baggrund, livserfaring, politisk overbevisning, bias og så videre. Hvis nævningen mener, at rehabilitering er ønskværdigt, er straffen langt mildere end, hvis nævningen mener, at afskrækkelse og fjernelse fra samfundet er målet. Disse bedømmelser handler dermed ikke om forbrydelsen og den anklagede, men om personlige præferencer.

På samme måde kan forskellige læreres personlige karakteristika afspejle forskellige bedømmelser af den samme danskopgave. Nogle lærere bedømmer den samme opgave hårdt, mens andre er generøse. Disse forskelligheder i bedømmelserne afspejler heller ikke retfærdighed, men beror på lærerens egne holdninger til karaktergivningspraksissen. Dermed er niveaustøj i karaktergivningen skadeligt for systemet, da lærerne ikke har den nødvendige konsensus i bedømmelserne, som det kræver, for at karaktergivningen beror på de samme forudsætninger og dermed er troværdig.

3.1.5 Mønsterstøj

Mængden af støj kommer an på kompleksiteten af problemstillingen. Hvis der er mere end én måde at opfatte en problemstilling på, så vil individer altid variere i deres vurderinger. Dog er alle bedømmere sikre på, at deres egen bedømmelse er den rigtige, da denne bedømmelse naturligt vil afskrive alternative udfald. Hvis individet kun kan se en løsning på problemstillingen, så vil individet tænke, at alle andre bedømmere selvfølgelig vil nå frem til samme løsning. Dømmekraftsproblemer defineres derfor som tvetydigheden i vurderinger forårsaget af modstridende præferencer blandt bedømmere (Kahneman et al., 2021: 202).

Hvis alle bedømmere var lige konsistente i deres hårdhed i bedømmelser (niveaustøjen), så ville det altid kunne forudsiges og udregnes, hvordan de ville bedømme i næste sag. Det er dog ikke sådan, det forholder sig. Niveauet af bedømmelserne er meget inkonsistente.

Et eksempel kan være, når en gruppe lærere bedømmer en danskopgave. Gennemsnitskarakteren for opgaven er 4. En tilfældigt udtrukket lærer har bedømt samme opgave til et 7-tal. Det er over gennemsnittet. Ved næste opgave, som har fået en gennemsnitlig vurdering på 7, burde den tilfældigt udtrukne lærer give karakteren 10, da denne lærer åbenbart er mere generøs end gennemsnittet. Men denne gang har læreren givet opgaven 4. Forklaringen er mønsterstøj (Kahneman et al., 2021: 74). Der findes to underliggende støjtyper i forbindelse med mønsterstøj.

Den første er stabil mønsterstøj. Denne type støj afspejler bedømmerens egen filosofiske forståelse af, hvad der er vigtigt at lægge vægt på i en bedømmelse. Her er det personlige præferencer, som kommer fra associationer. Præsenteres den tilfældigt udtrukne lærer for samme opgave året efter, må det forventes, at disse personlige præferencer dukker op igen. De er derfor kontinuerlige (Kahneman et al., 2021: 75). Måske går denne lærer meget op i, at der er korrekte kildehenvisninger i opgaven. Måske er læreren fuldstændig ligeglad med kvaliteten af kildehenvisninger. Disse associationer varierer fra individ til individ og er meget uforudsigelige. Dertil har individer også forskelligt vidensgrundlag forud for en bedømmelse. Associationerne stammer fra succeser, der har skabt selvtillid. Fejl, man gerne vil undgå at lave igen. Regler, som man husker godt, dem der er glemt og dem, der ignoreres. Alle disse associationer er unikke for hvert individ, da de kommer fra værdierne og personlige erfaringer. Dermed er individers stabile mønsterstøj også unik (Kahneman et al., 2021: 205).

Mønsterstøj indeholder også en anden type støj, som er situationsstøj, som er forbigående støj, der påvirker humøret. Vurderinger og bedømmelser afhænger utroligt meget af, hvordan humøret er for bedømmeren. Humøret ændrer, hvordan individer tænker. Hvis individet er i godt humør, vil det være mere positivt i sine vurderinger (Kahneman et al., 2021: 86-87). Det gode humør gør desuden individet mere samarbejdsvilligt og nemmere at forhandle med. Dog er det højst sandsynligt mere villig til at acceptere sin system-1-tænkning, hvilket betyder, at dets bias, førstehåndsindtryk og fordom påvirker dets tænkning (Kahneman et al., 2021: 87). Støjen kommer i det øjeblik, at individets humør påvirker dets dømmekraft, når det står over for komplekse vurderingsspørgsmål, som eksempelvis karaktergivning af danskopgaver. Tilgangen til problemet og dermed også konklusionen er afhængig af, om individet er træt,

stresset, tidspresset, om solen skinner eller om det regner. Og det er sjældent opmærksom på, at dets humør ændrer sig (Kahneman et al., 2021: 89). Selv rækkefølgen af vurderinger støjer for den næste vurdering, som skal foretages. Hvis en lærer eksempelvis har vurderet tre danskopgaver i træk til et 12-tal, så er der stor sandsynlighed for, at den fjerde danskopgave vil få en anden karakter, selvom den er berettiget til et 12-tal (Kahneman et al., 2021: 90).

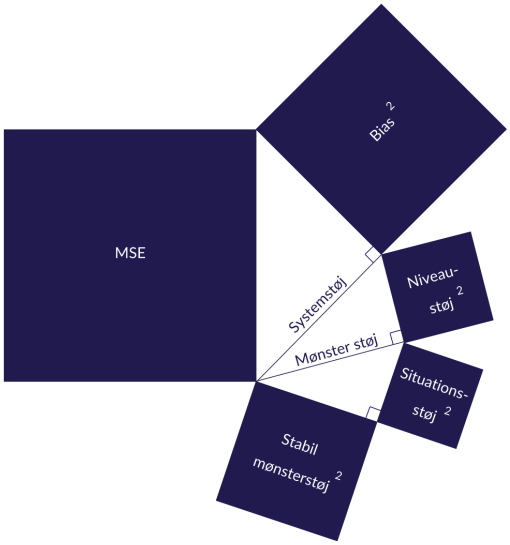
Det er svært at fjerne situationsstøj helt. I teorien kunne en kontrolleret setting reducere støj (Kahneman et al., 2021: 91). Dog kan de indre grunde til, at støjen opstår, aldrig styres. Kedsomhed, træthed, stress er alle faktorer, som ikke kan kontrolleres i et eksperiment, hvorfor sindet er et måleinstrument, der aldrig kommer til at måle perfekt (Kahneman et al., 2021: 93). Støjtyperne, der fører til fejl i bedømmelserne er skitseret i nedenstående tabel 3:

Støjtyper		
Type af støj	Definition	Eksempel
Bias	Psykologiske mekanismer, der fører til systematiske fejl i bedømmelser.	En gymnasielærer har en fordom om, at AI er en trussel mod ærlig og saglig viden og vil derfor på forhånd afvise mulige positive konsekvenser ved AI.
Niveaustøj	Variationen imellem bedømmeres personlige karakteristika.	Dommere i retssager dømmer den anklagede forskelligt afhængigt af, om dommeren mener rehabilitering er ønskværdigt, eller om afskrækkelse og fjernelse fra samfundet er målet.
Stabil mønsterstøj	Bedømmerens egen filosofiske forståelse af, hvad der er vigtigt at lægge vægt på i en bedømmelse.	En lærer vurderer kommatering, som det vigtigste i en opgave. En anden lærer vurderer refleksion som det vigtigste. Dermed lægger de vidt forskellige parametre til grund for deres samlede bedømmelse.
Situationsstøj	Forbigående støj, der påvirker humøret.	Læreren er stresset og træt. Det regner udenfor. Disse faktorer påvirker lærerens humør og dermed bedømmelsen negativt. Dette fungerer også modsat.

Tabel 3: Støjtyper - oversat frit efter Kahneman et al. (2021)

Det samlede fejlniveau i bedømmelsessituationer (senere kaldet *Mean Squared Error* (MSE)) kan efter gennemgangen af komponenterne inddeles i kategorier for overskuelighedens skyld. Fejl skyldes bias og systemstøj. Systemstøj kan inddeles i niveaustøj og mønsterstøj. Mønsterstøj kan inddeles i stabil mønsterstøj og situationsstøj (Kahneman et al., 2021: 211). Inden for mønsterstøjen er det den stabile mønsterstøj, der er dominerende i forhold til

situationsstøjen (Kahneman et al., 2021: 213). Dette er visualiseret ved følgende figur 2 (Kahneman et al., 2021: 211):



Figur 2: Fejl, bias og komponenterne i støj - oversat frit efter Kahneman et al. (2021): 211

3.1.6 Måder at reducere støj og bias

Kahneman et al. (2021) har forsket meget i, hvordan støj kan reduceres og bias kan fjernes i bedømmelser. Når man taler om *debiasing*, er der tre typer af løsninger:

Et eksempel er, at en badevægt konsekvent viser et halvt kilo for meget, når den bruges. Badevægten har derfor en systematisk bias, som der kan regnes med hver gang. Enten kan den viste vægt subtraheres med et halvt kilo hver gang. Denne metode kaldes *ex post*, hvor der justeres efter bedømmelsen er foretaget. Ellers kan det forsøges at justere badevægten, så den viser den rigtige vægt hver gang for eftertiden. Denne metode kaldes *ex ante*, hvor der interveneres før bedømmelsen foretages. Dermed sker der en præventiv *debiasing*. Kahneman et al. (2021) giver forskellige bud på, hvordan man kan intervenere før en bedømmelse (Kahneman et al., 2021: 236-237).

En metode er *boosting*. *Boosting* handler om at træne bedømmere til at være opmærksomme på deres bias og overkomme det. Man øger dermed bedømmeres kapaciteter, så de har mindre sandsynlighed for at lave statistiske fejl (Kahneman et al., 2021: 238). Generelt er træning, uddannelse og intelligens betydningsfulde faktorer, der alle sammen påvirker niveauet af bedømmelser. Intelligens påvirker jobperformance positivt og sammenkædes med højere

akademisk niveau (Kahneman et al., 2021: 228-229). Dette vil højst sandsynligt kunne ses afspejlet i fejlniveauet for de lærere, der *er* eller *har været* censor. Som nævnt i problemfeltet får disse gymnasielærere vejledning af fagkonsulenten og deltager desuden i kurser, der skal sikre, at bedømmelserne bliver så korrekte og nøjagtige som muligt (Styrelsen for Undervisning og Kvalitet, 2024: 1). Disse lærere har dermed bedre forudsætninger for at lave færre fejl. Denne del af teorien er basis for projektets underhypotese. Den primære hypotese (H1a) vil blive præsenteret senere i afsnit 3.1.11. Underhypotesen lyder som følger:

H1b: Gymnasielærere, der er eller har været censorer og som anvender beslutningsstøtteværktøjet, vil have et lavere samlet fejlniveau i forhold til de lærere, der ikke har censorerfaring.

En anden metode er *nudging*. *Nudging* foregår ved at modificere miljøet, hvor bedømmelsen bliver foretaget. Ved at modificere kan man reducere eller tiltrække bias i en given retning for at gøre bedømmelsen mere korrekt. Det kunne eksempelvis være, at gøre den rigtige beslutning til den nemmeste beslutning (Kahneman et al., 2021: 237). Ulempen ved alle strategierne, som Kahneman et al. (2021) oplister, er, at det foregives, at der er bias til stede i beslutningsprocessen. Det er oftest ikke sådan, som det regnes med, at det er (Kahneman et al., 2021: 239).

Bias er fejl, som vi kan observere og måle. De er retningsbestemte og kan reduceres ved simpel *nudging* og andre interventioner. Dog er støj, som tidligere nævnt, uforudsigelige fejl, der hverken kan ses eller forklares. Alligevel er disse fejl ofte negligeret, selvom disse fejl er mindst lige så skadelige for beslutningsprocessen som bias (Kahneman et al., 2021: 243). Kahneman et al. (2021) sammenligner støjreduktion med håndvask; der fjernes skidt og uønskede bakterier, men det vides ikke hvilke. På samme måde er støjreducerende teknikker i beslutningsprocessen statistisk set med til at fjerne støj, uden det er klart hvilke fejl, der forebygges imod (Kahneman et al., 2021: 243-244).

3.1.7 Social influence

I teorien forholder det sig sådan, at når der stilles en gruppe uafhængige individer et spørgsmål, så er sandsynligheden for, at gennemsnittet rammer plet, stor (Kahneman et al., 2021: 99). Dette kaldes *Wisdom-of-crowds*-tesen. Dermed burde gennemsnitskarakteren for den samme

opgave være den mest korrekte. Det eneste, der dog kan forpurre denne tese er, hvis besvarelserne *ikke* er uafhængige. Dette kaldes *social influence*, som kan underminere *wisdom-of-crowds* (Kahneman et al., 2021: 99).

Social influence skaber en høj tilstedeværelse af støj. Mennesker bliver påvirket af mennesker. Den første, der siger sin holdning i en gruppe af mennesker, påvirker de resterende gruppemedlemmer så meget, at de højst sandsynligt adopterer denne holdning (Kahneman et al., 2021: 96-97). Denne effekt er så kraftfuld, at den kan producere store skift i holdningsdannelsen for andre (Kahneman et al., 2021: 98). Denne tese læner sig ydermere op af konklusionsbias, hvor individer accepterer system-1-tænkningens hurtige og intuitive løsningsforslag (Kahneman et al., 2021: 169).

3.1.8 Biasing information

En anden måde at reducere støj i beslutningsprocessen er ved at nudge i en given retning ved at påvirke beslutningstageren med *biasing information*. Som eksempel henviser Kahneman et al. (2021) til en undersøgelse, hvor teknikere skulle sammenligne fingeraftryk, som de plejer i afdelingen for opklarende politiarbejde. En metode, som aldrig har været mistro for indtil 2002. Teknikeren finder matchet af fingeraftryk, og dette bruges som bevisførsel mod den anklagede. År senere bliver teknikeren præsenteret for de samme fingeraftryk, hvortil der bliver tilføjet information om sagen. "Den anklagede har tilstået." "Øjenvidnet er sikker." Eller i den modsatte ende: "Den anklagede har et alibi." "Våbnet passer ikke med patronerne på gerningsstedet." (Kahneman et al., 2021: 249). Afhængigt af, hvilken information teknikeren fik, ville teknikeren ubevidst ændre den oprindelige bedømmelse. Teknikerne, som fik suppleret kontekstuel information, aktiverer konklusionsbias i en given retning. På den måde er der tale om *ex ante* og *nudging*, da denne *biasing information* intervenerer før bedømmelsen tages (Kahneman et al., 2021: 250). Det er plausibelt at antage, at danske gymnasielærere, som modtager *biasing information,* inden de giver feedback og karakterer på en danskopgave, vil ændre deres vurdering og aktivere konklusionsbias. Denne information kunne være "Den her opgave har allerede fået et 7-tal af en anden lærer." På den måde vil flere lærere være enige. *Biasing information* ændrer, *hvad* bedømmerne opfatter, foruden *hvordan* de opfatter det (Kahneman et al., 2021: 250).

3.1.9 Second opinion

I forlængelse af denne metode er det naturligt at uddybe netop den *biasing information*, de bliver præsenteret for. En anden fagprofessionel kan foretage en *second opinion* i virkeligheden. I sundhedsverdenen foretager lægefagligt personale vurderinger hele tiden. Det er ikke svært at diagnosticere en skulder, der er gået af led, men i mere tekniske situationer er det ofte et krav på hospitaler, at en anden læge, uafhængigt fra første læge, også skal se på patienten (Kahneman et al., 2021: 274). I mere tekniske situationer gør læger brug af beregninger for at kunne diagnosticere patienter. Disse beregninger erstatter nogle gange menneskelig dømmekraft helt. Eksempelvis ved bakterieskrab, der testes og kommer retur med et testresultat, som lægen diagnosticerer efter (Kahneman et al., 2021: 274).

På hospitalerne benytter man sig, ifølge Kahneman et al. (2021) af tre forskellige støjreducerende teknikker for at sikre bedre medicinske bedømmelser. Træning af personale (*boosting*), sammenlægning af flere ekspertvurderinger (*second opinion*) og brug af algoritmer og kunstig intelligens. Ifølge Kahneman et al. (2021) vil medicinske beslutninger i større grad være baseret på algoritmer og AI. AI reducerer bias og støj, hvilket sparer hospitalet penge og redder liv (Kahneman et al., 2021: 280). Men hvad er AI egentlig og fungerer det i beslutningssituationer?

3.1.10 Algoritmer og AI

Artificial intelligence, også kaldet AI, udvikler sig, som nævnt tidligere i opgaven, med hastige skridt. AI overtager opgaver, som man tidligere har betragtet som udelukkende opgaver, mennesker kunne håndtere. I dag kan AI give præcis rutevejledning, genkende ansigter, oversætte sprog, generere billeder og tekst, som et menneske kunne have gjort (Kahneman et al., 2021: 123). Alt sammen på grund af avancerede computere, sprogmodeller og algoritmer, der langsomt udkonkurrerer menneskelig dømmekraft, fordi disse er støjløse (Kahneman et al., 2021: 124). Dog afhænger kvaliteten af den kunstige intelligens af, hvor stort et datasæt der kan trækkes på. Sofistikerede analyser kræver store datasæt, og de er i hastig udvikling (Kahneman et al., 2021: 129). Ved brug af mønsterfinding kan algoritmerne forudsige udfald på baggrund af det data, som algoritmen har fået inkorporeret på forhånd (Kahneman et al., 2021: 130). Derfor bruges disse hyppigere i beslutningssammenhænge, fordi menneskeligt skøn ikke er præcist nok i forhold til AI, som kan øge nøjagtigheden af vurderinger markant (Kahneman et al., 2021: 131-132).

3.1.11 AI som beslutningsstøtteværktøj

Argumenterne for ikke at tage støjreducerende teknikker i brug på arbejdspladserne er mange. Det er for dyrt at implementere, og det er for besværligt at omstille sig (Kahneman et al., 2021: 329). I praksis støjer en lærer rigtig meget, når det normeres efter at give en karakter hver halve time (UVM, 2023d). For at undgå dette kan en kollega vurdere den samme opgave, hvilket i praksis fungerer som en *second opinion*. Læreren kan bruge længere tid på at gennemgå opgaven eller følge en detaljeret checkliste i forbindelse med vurderingen (Kahneman et al., 2021: 329). Alternativt kan læreren anvende kunstig intelligens som et beslutningsstøtteværktøj. AI er støjløs og uden bias (Kahneman et al., 2021: 334). Dermed vil læreren kunne reducere en del af niveaustøjen, den stabile mønsterstøj og lærerens bias, som opstår i alle bedømmelsessituationer. Denne løsning vil kunne forbedre den menneskelige dømmekraft i situationer, hvor personlige præferencer, subjektivitet, individualitet og kreativitet ikke er ønskværdigt. Passende brug af AI i en karaktergivningssituation gør beslutningen mindre afhængig af én fagprofessionel og dermed mere objektiv og tættere på den sande værdi (Kahneman et al., 2021: 371). Dette leder op til projektets primære hypotese:

H1a: Gymnasielærere, som har benyttet AI som beslutningsstøtteværktøj, vil i gennemsnit lave færre fejl i karaktergivningen i forhold til kontrolgruppen.

3.1.12 Objektiv ignorance

Ifølge Kahneman et al. (2021) findes der ingen grunde til, at man ikke burde udnytte potentialet i AI langt mere i bedømmelsessituationer, end det er tilfældet i dag. Alligevel ignoreres faktummet, at AI vurderer bedre end mennesker (Kahneman et al., 2021: 134). Eksperter stoler på sin egen dømmekraft og peger i stedet på, at AI dehumaniserer beslutninger. Det er et problem, hvis ansvar fjernes fra mennesker, medarbejdere og ledere, for hvem har så ansvaret? Svaret på, hvorfor eksperter og fagprofessionelle ikke stoler på AI, er ifølge Kahneman et al. (2021) en blanding af sociopsykologiske faktorer som "frygten for teknologisk arbejdsløshed", "dårlig oplysning" og "generel utilfredshed over computere" (Kahneman et al., 2021: 134).
Ifølge Kahneman et al. (2021) er objektiv ignorance årsagen til, at menneskelig dømmekraft aldrig bliver erstattet af AI. Eksperter og fagprofessionelle vil have et perfekt alternativ til menneskelig dømmekraft med en succesrate på 100 procent. Men det er umuligt (Kahneman et al., 2021: 145-146). I realiteten findes der ingen opgaver, hvor menneskelig dømmekraft er signifikant dårligere, og hvor algoritmer er markant bedre, hvor den samme mængde

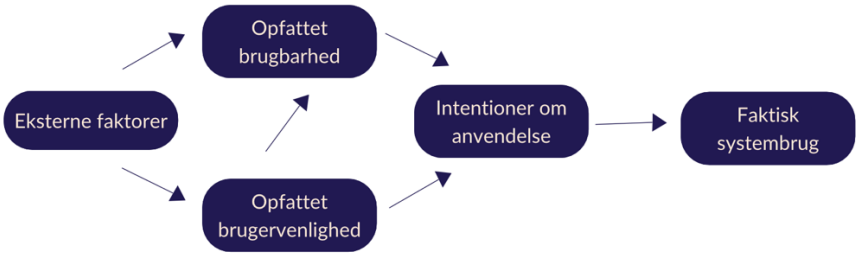
information er givet (Kahneman et al., 2021: 143). Dette er en af grundene til, at beslutningstagere højest sandsynligt vil fornægte at overlade beslutninger til AI, hvorefter intuitionen og dømmekraften forsvinder (Kahneman et al., 2021: 146).

Kahneman et al. (2021) mener, at debatten er misforstået. AI kommer aldrig til at erstatte menneskelig dømmekraft - men den kan forbedre den (Kahneman et al., 2021: 134). Kahneman et al. (2021) præsenterer forskning, der viser, at almindelige mennesker ikke er mistroiske over for AI. Når mennesker skal vælge mellem AI eller et menneske til at rådgive dem, så falder svaret ofte på AI. De giver AI en chance, men så snart de laver fejl, falder tilliden markant. Mennesker forventer, at AI foretager perfekte vurderinger. Derfor ser de ingen grund til at bruge AI, hvis ikke den er 100 procent perfekt (Kahneman et al., 2021: 135). Generelt er det svært for mennesker at ændre deres adfærd, medmindre alternativet er nær-perfekt. Men hvad ville være det foretrukne valg, hvis der er mulighed for at købe et skrabelod enten med 59 procent chance for gevinst eller 65 procent chance for gevinst til samme pris (Kahneman et al., 2021: 136). Derfor stiller Kahneman et al. (2021) spørgsmålet: "*AI laver fejl. Selvfølgelig gør den det. Men hvis menneskelig dømmekraft laver flere fejl, hvilken af delene skal vi så stole på?*" (Kahneman et al., 2021: 136).

3.2 Teknologi Accept Modellen

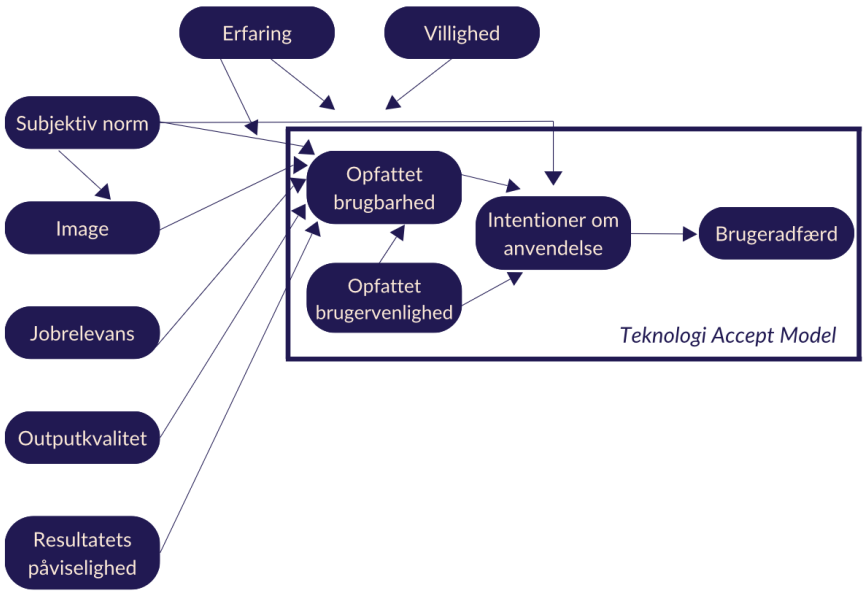
Digitaliseringen i Danmark har i løbet af de seneste årtier udviklet sig hurtigt og er med tiden blevet et alment værktøj i mange danskeres hverdag. Introduktionen af digitale offentlige tjenester som hjemmesiden borger.dk, MitID og det digitale sundhedskort har gjort det muligt for borgerne at komme i direkte kontakt med det offentlige og deres selvbetjeningsmuligheder (Jacobsen et al., 2023).

For at ny informationsteknologi skal kunne implementeres og forbedre produktiviteten, skal den først accepteres af brugeren. For at undersøge dette anvendes den såkaldte Teknologi Accept Model (TAM). Modellen, som blev introduceret i 1989 og endeligt uddybet af Davis og Venkatesh i 1996, er baseret på teorien om begrundet handling. TAM har til formål at skitsere, hvordan individer eller grupper accepterer og anvender teknologi. Dertil også hvilke eksterne faktorer, der har en afgørende effekt på personens attitude, mod og intentioner om at anvende og implementere den nye teknologi i sin hverdag (Chuttur, 2009: 1-10). Modellen ses skitseret nedenfor i figur 3:



Figur 3: Teknologi Accept Modellen - oversat frit efter Venkatesh & Davis (1996)

Som videreudvikling på den originale model præsenterede Venkatesh og Davis i 2000 TAM2. Heri præciserede de den del af modellen angående eksterne faktorer og præsenterede de enkelte punkter, der kunne tænkes at have indflydelse på individets accept af teknologi (Venkatesh & Davis, 2000: 186-187). Senere har de videreudviklet på TAM, men det er denne udgave som projektet vil tage udgangspunkt i. Denne udgave af modellen ses skitseret nedenfor i figur 4.



Figur 4: Teknologi Accept Modellen 2 - oversat frit efter Venkatesh & Davis (2000)

Denne udgave af TAM præsenterer de syv faktorer, der har betydning for, hvorvidt den præsenterede teknologi kan accepteres af individet eller gruppen. Heri kategoriseres de i to grupperinger. Henholdsvis den sociale indflydelse (subjektive normer, villighed, og image) samt den kognitive proces (jobrelevans, kvaliteten af teknologiens produkt og resultatets påviselighed). Ydermere inkorporeres også individets egen erfaring med det teknologiske system (Venkatesh & Davis, 2000: 187-189).

Den subjektive norm er baseret på modellens udspring fra teorien om begrundet handling. Det defineres i litteraturen som: *"individets opfattelse af, hvorvidt de fleste personer i deres omgangskreds, hvis mening betyder noget, synes at individet skal underlægge sig den anbefalede teknologi"*. Her skal det forstås som, at individet vægter andres meninger om en selv højere end sin egen holdning til eksempelvis integreringen af nye IT-systemer. Villighed refererer til den grad individet oplever beslutningen omhandlende integreringen som obligatorisk, samt hvad straffen for ikke at følge den ønskede udvikling er. Image omhandler, hvorledes implementeringen og den eventuelle accept af teknologien har en indvirkning på individets sociale status (Venkatesh & Davis, 2000: 187-190).

I forbindelse med den kognitive proces vurderer individet først teknologiens relevans for sit job. For at opnå en tilfredshed med teknologien skal der kunne ses tydelige sammenhæng med produktet og arbejdsopgaverne. Samtidig vurderes kvaliteten af teknologiens output også. Den endelige vurdering af brugbarheden skal kunne ses i sammenligning med arbejdet. Individet skal kunne stole på og være enig i det arbejde, som teknologien producerer. Slutteligt skal resultaterne fremgå som brugbare og vurderes anvendelige i symbiose med arbejdet (Venkatesh & Davis, 2000: 190-192).

Disse faktorer fra henholdsvis den sociale indflydelse samt den kognitive proces vil have en direkte effekt på den opfattede brugbarhed af den teknologi, der ønskes accepteret. Heraf også hvilke formål det ønskes at anvende teknologien til, samt hvor nem og lige til teknologien er at anvende, hvilket slutteligt vil kunne tegne et billede af, hvordan individet vil anvende og acceptere den teknologi. Slutteligt er effekten af erfaring med det produkt som ønskes implementeret. Her skal det forstås, at et individ med en større grad af erfaring enten med lignende teknologi - eller blot teknologi generelt - vil have svagere negativ indvirkning på deres vurdering af produktets brugbarhed (Venkatesh & Davis, 2000: 189-190). Dette leder op til projektets sidste hypotese H2, der lyder som følger:

H2: Eksperimentgruppen AI bias vil være mere tilbøjelig til at være uenige med feedback- og karaktergivning genereret af AI i forhold til eksperimentgruppen Second Opinion Influence.

Dermed vil dette projekts teoretiske apparat baseres på Kahneman et al.s (2021) teori om støj og Venkatesh og Davis' (2000) teori om Teknologi Accept Modellen. Herunder vil projektet primært gøre brug af begreber som bias, støj, *boosting*, *nudging*, konklusionsbias, *second opinion*, *biasing information*, *social influence*, objektiv ignorance, opfattet brugbarhed, outputkvalitet, den subjektive

norm og jobrelevans for til sidst at tolke på gymnasielæreres brugeradfærd ved brugen af AI som beslutningsstøtteværktøj.

4. Metode

I dette afsnit vil de metodiske overvejelser, som ligger til grund for projektet, gennemgås. Formålet er at belyse de valg, der er truffet i forhold til dataindsamling, analysemetoder og forskningsdesign for at højne transparensen af undersøgelsen.
Først og fremmest vil projektets videnskabsteoretiske position blive præsenteret og begrundet i følgende afsnit.

4.1 Kritisk realisme

Dette projekt vil tage udgangspunkt i en realistisk position inden for videnskabsteori. Herunder synes både støj-teorien (Kahneman et al., 2021) og TAM (Venkatesh & Davis, 2000) at lægge sig op af kritisk realisme.

Den kritiske realismes ontologi tager udgangspunkt i, at verden eksisterer uafhængigt af observatøren (Ingemann, 2020: 88). Desuden anses virkeligheden for kompleks og kontekstuel. Disse forbehold strækker sig helt tilbage til Schmollers standpunkt i *Methodenstreit* (Ingemann, 2020: 88-89). I dette projekt kommer disse forbehold til syne i form af de faktorer som ikke er målbare igennem de udvalgte teorier. Erkendelsen af et samfunds mange lag og diverse indvirkninger vil have en tydelig effekt på resultaterne af undersøgelsen. Med afsæt i den kritiske realisme må virkeligheden anerkendes som kontekstuel. Den afhænger derfor af genstandsfeltet selv og de historiske betingelser. Hvervet som gymnasielærer og denne del af den offentlige sektor har udviklet sig meget i de seneste år. Dette forudsætter da en anerkendelse af, at denne udvikling er konstant (Ingemann, 2020: 89).

Ved at påtage sig den kritiske realisme som videnskabsteoretisk overbevisning vil de ontologiske og epistemologiske forbehold skabe problematikker for hinanden. Når virkeligheden er kontekstuel, vil den være svær at gennemskue. Samtidig er det svært at finde de generelle lovmæssigheder. Derfor har dette projekt på baggrund af de observationer, der er lavet af processen for karaktergivning, deduceret og beskrevet det samfundsvidenskabelige problem (Ingemann, 2020: 90-91). Som forlængelse af denne procedure vil det være muligt, på baggrund af operationaliseringen af den udvalgte teori, at systematisere de mønstre som undersøgelsen vil finde. Disse vil gennem den kvantitative metode kunne præsenteres igennem grafikker og hypotesetests. Derved vil projektet have induceret det observerede problem og systematiseret det umiddelbare ved at beskrive de mønstre, undersøgelsen belyser (Ingemann,

2020: 91). Projektet anvender ydermere kvalitativ metode, da der også ønskes en forståelse af konteksten til hvorfor gymnasielærere støjer og er under indflydelse af bias. Videnskabsteorien anerkender samtidig vigtigheden af kontekst for den sande forklaring. Den kontekst, hvor individet eksisterer, påvirker og socialiserer mennesket, og styrer dermed dets tanker og handlinger. Dog kan denne påvirkning fejle, og individet har da potentialet til at forandre strukturen og konteksten (Ingemann, 2020: 95).

Der opstår dog et problem, når det kommer til at forklare det, som undersøgelsen har beskrevet og observeret. Ifølge den kritiske realisme er forklaringen ikke observerbar, da den ligger skjult (Ingemann, 2020: 91). Derfor præsenterer den kritiske realisme en anden fremgangsmåde end fra de andre videnskabsteorier. Ingemann (2020) beskriver metoden som et isbjerg, hvor det ummiddelbare er toppen, men der er en stor del af bjerget, som er under vandoverfladen, og derfor ikke er sanselig - og man må derfor dykke ned for at se hele bjerget (Ingemann, 2020: 91-92). Derfor anvender man i den kritiske realisme en mere rationalistisk tilgang og tager sine forbehold ud fra den udvalgte teori. Ved at tage sit afsæt i teorier om genstanden man vil observere og forklare, vil det være muligt at opnå en mere gennemgående forklaring af observationerne baseret på de teoretiske forestillinger (Ingemann, 2020: 92).

Ønsket med denne undersøgelse og kritisk realisme som valg af videnskabsteoretisk overbevisning er, at processen skal ende i det dybe stratum. (Ingemann, 2020: 92). Det ønskes altså at afdække de kausale sammenhænge, der er årsag til undersøgelsens problemstilling. Her vil undersøgelsens teorier igen have en betydelig effekt. Som beskrevet anerkender den kritiske realisme virkelighedens mange lag - dertil også de uforklarlige samt de biologiske indvirkninger. Kahneman et al.s (2021) teori om støj forudsætter, at de danske gymnasielærere ikke altid er i fuld kontrol over deres endelige bedømmelse. TAM bevidner, at individet ikke selv er enerådig om, hvorvidt en implementering af en kunstig intelligens som beslutningsstøtteværktøj vil være acceptabel i deres hverv. Forklaringerne på denne undersøgelses problemformulering må derfor anerkendes som multikausale, fordi de eksisterer i samspil mellem forskellige, kausale potentialer, der kan befinde sig på flere niveauer af virkelighedens lag (Ingemann, 2020: 93-94).

4.2 Valg af forskningsdesign

Formålet med dette afsnit er at klarlægge det forskningsdesign, som ligger til grund for projektets undersøgelse.

Det eksperimentelle forskningsdesign anvendes sjældent i sin sande form inden for samfundsvidenskab. Men i sammenhænge, hvor formålet er at finde en kausal sammenhæng mellem variable, er dette design højt værdsat, da det skaber meget troværdige resultater med en høj grad af intern validitet (Bryman, 2016: 44). For at denne type design skal kunne fungere i praksis, er det nødvendigt at manipulere med problemformuleringens uafhængige variabel. For denne undersøgelse er den uafhængige variabel AI som beslutningsstøtteværktøj. Ved at manipulere variablen vil det være muligt for projektet at konkludere, hvorvidt denne har en afgørende effekt på de grupperinger af deltagere, som projektet har lavet. Udformningen af projektets forskningsdesign vil bestå af et såkaldt randomiseret kontrolleret forsøg (RCT). Dette er det klassiske design for eksperimenter, hvori der laves tilfældige grupper, som hver især bliver udsat for variationer af den manipulerede variabel. Dette design bliver også anerkendt som 'guldstandarden', som andre undersøgelser holdes op imod (Bryman, 2016: 44-45). Ydermere er tilstedeværelsen af undersøgelsens kontrolgruppe med til at fjerne de andre forklaringer på det observerede fænomen (Bryman, 2016: 46).

4.3 Valg af metode

Til at besvare projektets opstillede problemformulering inden for det valgte forskningsdesign samt videnskabsteoretiske perspektiv, vurderes en kombination af kvantitativ- og kvalitativ metode som værende mest fordelagtig. Dette vil ske gennem statistisk databehandling samt inddragelse af kvalitative udsagn i projektets analyse.

Kvantitativ metode skaber med sin deskriptive tilgang til databehandling et større overblik over og indsigt i udvalgte variable. Projektets data, som behandles ved kvantitativ metode i analysen, findes gennem projektets surveyeksperiment.

Desuden benyttes også kvalitativ metode i projektets undersøgelse af problemstillingen. For at redegøre for, at gymnasielærernes feedback er modstridende og utilfredsstillende, vil der blive gjort brug af kvalitativ metode. Gennem udsagn fra gymnasielærernes angivne feedback vil det afdækkes, at dette foregår i undersøgelsens stikprøve. Her skal den kvalitative metode bidrage med en dybere forståelse af de kvantitative fund, hvorved projektet nærmer sig det dybe stratum. Dog vil der i projektets primære hypotese (H1a) og underhypotese (H1b) udelukkende blive gjort brug af kvantitativ metode, mens der i projektets anden hypotese (H2) vil blive gjort brug af både kvantitativ- og kvalitativ metode. Herved nærmer projektets undersøgelse sig mixed methods-forskning, hvor der ved brug af triangulering tilstræbes konvergent validering (Frederiksen i Brinkmann & Tanggaard, 2020: 261-262). Det vil sige, at

undersøgelsens konkluderende resultater som udgangspunkt skal bakkes op af begge metoder. På den måde vil undersøgelsens resultater være stærkere underbygget.

4.4 Surveyeksperiment

Efter at have præsenteret og argumenteret for valg af videnskabsteoretisk ståsted, forskningsdesign og metode, vil der i det følgende blive redegjort for udførelsen af projektets surveyeksperiment (bilag 1), der danner al data, som benyttes i analysen.

For at styrke undersøgelsens økologiske validitet blev surveyen sendt til en forhenværende gymnasielærer i dansk forud for distribuering til målgruppen. Hensigten med dette var at få input fra en fagprofessionel, så oplevelsen for gymnasielærerne blev bedre og desuden mere virkelighedsnær og gyldig. Dertil gav den forhenværende gymnasielærer i dansk indblik i kognitive tendenser, som projektgruppen ikke ville have mulighed for selv at nå frem til. De fagprofessionelle input vil blive uddybet yderligere i de relevante afsnit nedenfor.

4.4.1 Kriterier for deltagere og distribuering

Projektets surveyeksperiment har til formål at undersøge bedømmelsespraksissen i skriftlig dansk i gymnasiet på A-niveau. Målgruppen for projektets surveyeksperiment er derfor udelukkende gymnasielærere, som underviser i dansk på A-niveau. Med henblik på at udelukkende distribuere surveyen til målgruppen, er der foretaget en systematisk gennemgang af alle danske gymnasiers hjemmesider på baggrund af en oversigt fra Danske Gymnasier (2024). På hvert enkelt gymnasiums hjemmeside er der taget udgangspunkt i gymnasiernes liste over undervisere, hvor mailadresserne på de, der figurerer som undervisere i dansk, er tilføjet til distributionsmaillisten. Efter endt indsamling indbefatter denne liste omkring 2000 respondenter, som alle oprettes i SurveyXact, der er et system til spørgeskemaundersøgelser. Ved at have oprettet alle gymnasielærere i dansk som respondenter i SurveyXact, findes det nemt og overskueligt at udsende distributions- og rykkermails samt være i dialog med de gymnasielærere, der ønsker det.

Følgende distributionsmail i figur 5 blev udsendt til cirka 2000 gymnasielærere, som underviser i dansk:

Hejsa! Vi har brug for DIN hjælp!

Vi er en studiegruppe fra Aalborg Universitet, som er i gang med at udarbejde vores afsluttende bachelorprojekt i Politik & Administration. Vores problemstilling udspringer af det pres, som vi ved, at gymnasielærere undergår i perioder. Vi vil i den forbindelse undersøge kompleksiteten af bedømmelsespraksissen i dansk A.

I den forbindelse håber vi virkelig, at du vil afsætte noget af din sparsomme tid til at deltage i vores undersøgelse. Undersøgelsen indebærer din vurdering af en kort besvarelse til en eksamensopgave i dansk på A-niveau.

Vi ved, at det er en ubelejlig periode, vi skriver i, så du skal vide, at dit bidrag vil blive påskønnet i meget høj grad! Du er desuden mere end velkommen til at dele undersøgelsen med andre gymnasielærere i dansk i dit netværk.

Du vil naturligvis fremstå fuldkommen anonym.

Undersøgelsen kan tilgås via følgende link:
<%MorpheusMailLink%>

På forhånd tusind tak!

Mange håbefulde hilsner,
Adam Alexander Faust Spies
Rasmus Jakobsen
Maria Rom Kristiansen

Figur 5: Distributionsmail til gymnasielærere

Denne fremgangsmåde for distribution vurderes effektiv, da respondenterne fik tilsendt undersøgelsen direkte i deres mailindbakke. Desuden kan respondenterne nemt tilgås og behandles via SurveyXact.

For bedst muligt at sikre, at respondenterne modtog mailen og deltog i spørgeskemaet, blev der også distribueret en rykkermail en uge efter den originale mail. Følgende rykkermail i figur 6 blev sendt til dem, som ikke havde svaret på spørgeskemaet efter første runde:


Hejsa!

Vi er en studiegruppe fra Aalborg Universitet, som kontaktede dig i sidste uge, da vi er ved at udarbejde vores afsluttende bachelorprojekt.

Vi søger gymnasielærere i dansk på A-niveau til besvarelse af vores spørgeskema.
Vi ønsker at undersøge, hvorvidt det er muligt at lempe det pres, som gymnasielærere oplever i deres arbejde med feedback- og karaktergivning af skriftlige opgaver og derved kunne lette arbejdsbyrden.

Vi ved, at det er en ubelejlig periode, vi skriver i, så du skal vide, at dit bidrag vil blive påskønnet i meget høj grad! Du er desuden mere end velkommen til at dele undersøgelsen med andre gymnasielærere i dansk i dit netværk.

Du vil gennem hele besvarelsen forekomme anonym.

Spørgeskemaet kan tilgås via følgende link:

<%MorpheusMailLink%>

HUSK at trykke "afslut", når du er igennem alle spørgsmål i spørgeskemaet!

På forhånd tusind tak!

Mange håbefulde hilsner,
Adam Alexander Faust Spies
Rasmus Jakobsen
Maria Rom Kristiansen

Figur 6: Rykkermail til gymnasielærere

Distributionen varede præcis to uger inklusiv distributions- og rykkermail.
Efter endt distribution har i alt 461 gymnasielærere besvaret surveyen, hvoraf 119 har gennemført hele surveyen. I perioden med distribuering modtog projektgruppen en del henvendelser om, at flere gymnasielærere havde intentioner om at besvare spørgeskemaet. Men grundet surveyens omfang, og at denne blev udsendt i en ubelejlig periode kortvarigt inden diverse eksamener, blev surveyen nedprioriteret af flere. Dette anses naturligvis for ærgerligt for undersøgelsens repræsentativitet og eksterne validitet, men dette er desuden med til at understrege, at projektets problemstilling omkring gymnasielærernes arbejdspres er reel.

Ved distribution i SurveyXact har respondenterne mulighed for at besvare den udsendte mail. Denne mulighed for at sende en returmail var der over 100 gymnasielærere, der benyttede sig af. På baggrund af dette indledte der sig mange korrespondancer med gymnasielærere, som omhandlede blandt andet arbejdspres, holdning til kunstig intelligens i undervisningen og

holdninger til projektets bidrag. Disse anses substantielle for undersøgelsen af projektets problemformulering, da de er med til at rammesætte problemstillingen. Projektgruppen har via disse korrespondancer fået øje for, hvad der er essentielt i den samfundsmæssige problemstilling, hvorfor dette øger projektets økologiske validitet.
Disse rammesættende input vil indgå i det videre arbejde med problemstillingen i projektets analyse.

4.4.2 Inddeling i kontrol- og eksperimentgrupper

Som beskrevet i afsnit 4.2 omkring valg af forskningsdesign, anvender dette projekt et RCT-eksperiment design. Det fordrer inddeling af respondenterne i kontrol- og eksperimentgrupper. Det ønskes undersøgt, hvorvidt brugen af AI som beslutningsstøtteværktøj højner korrektheden i karaktergivningen. Desuden ønskes det undersøgt, hvorvidt visheden om, at beslutningsstøtteværktøjet består af feedback samt karakter angivet af kunstig intelligens, har en effekt på gymnasielærernes holdning til beslutningsstøtteværktøjet.

Gennem projektets surveyeksperiment inddeles respondenterne tilfældigt i tre grupper på baggrund af fødselsmåned. To af dem vil agere eksperimentgrupper, og den sidste vil fungere som kontrolgruppe. Formålet med denne inddeling er at teste projektets teoriapparat. Undersøgelsens kontrolgruppe vil kunne bistå begge eksperimentgruppers fund ved at fjerne alternative forklaringer på problemet, de hver især undersøger (Bryman, 2016: 46). Dermed kan der fokuseres på den eller de kausale sammenhænge, som projektets analyse finder, mens den samlede undersøgelses interne validitet vil kunne styrkes.

I et RCT-eksperiment inddeles deltagerne normalt i to grupper, henholdsvis en eksperimentgruppe og en kontrolgruppe (Bryman, 2016: 45). Men da denne undersøgelse skal redegøre for flere teorier, som begge har en effekt på problemformuleringens uafhængige variabel, inddeles deltagerne i tre grupper. To af disse tre grupper, vil da udsættes for en intervention af den manipulerede uafhængige variabel, hvorfra deres effekt på den afhængige variabel, som er korrektheden af feedback- og karaktergivningen, vil blive målt og analyseret. Som nævnt inddeles alle respondenter tilfældigt i de tre grupper på baggrund af deres fødselsdagsmåned. Respondenter, som har fødselsdag i januar, februar, marts eller april, vil indgå i den første eksperimentgruppe, som ikke ved, at beslutningsstøtteværktøjet er genereret af kunstig intelligens. Denne kaldes *second opinion influence* (herefter SOI), fordi disse, i henhold til Kahneman et al.s (2021) teori om støj, oplever feedbacken og karakteren som en

second opinion. Denne *second opinion* burde aktivere konklusionsbias som følge af tesen om social indflydelse, der påvirker individers system-1-tænkning i en given retning.

Respondenter, som har fødselsdag i maj, juni, juli eller august, vil indgå i den anden eksperimentgruppe, som bliver oplyst om, at beslutningsstøtteværktøjet er genereret af kunstig intelligens. Denne gruppe kaldes AI Bias på baggrund af Kahneman et al.s (2021) tese om objektiv ignorance. Denne ignorance resulterer i en negativ bias om kunstig intelligens for fagprofessionelle, der normalt bruger deres fagprofessionelle skøn i forbindelse med bedømmelser.

Den sidste gruppe vil agere kontrolgruppe uden intervention, og udgøres af respondenter, som har fødselsdag i september, oktober, november eller december.

De respektive gruppers inddeling samt respondenternes fordeling er skitseret i nedenstående tabel 4:

Fordeling af respondenter i surveyeksperimentets tre grupper			
	Second opnion influence	AI bias	Kontrolgruppe
Måned	Januar, februar, marts og april	Maj, juni, juli og august	September, oktober, november og december
Antal	37	49	33

Tabel 4: Fordeling af respondenter i tre grupper

4.4.3 Opgavebeskrivelsen og -besvarelsen

Ved deltagelse i surveyeksperimentet vil alle respondenter blive præsenteret for undersøgelsens opgavebeskrivelse samt besvarelsen fra en 3.g elev. Alle respondenter vil ligeledes blive bedt om at angive en samlet karakter samt feedback til besvarelsen.

Opgavebeskrivelsen er fra stx skriftlig dansk eksamen fra den 23. maj 2023. Projektgruppen har fået aktindsigt til dette prøvesæt fra Styrelsen for undervisning og kvalitet: kontor for

prøver, eksamen og test. På grund af brud på ophavsret er det ikke muligt at benytte den eksakte opgavebeskrivelse fra eksamenssættet, hvorfor den er omskrevet således, at den lægger sig ganske tæt op ad den oprindelige. Denne ses nedenfor i figur 7:

Figur 7: Omskrevet opgaveformulering: Skriftlig dansk, stx, 23. maj 2023

Som det fremgår af opgaveformuleringen, er der fire vedlagte tekster: Primærteksten "Er der noget dansk litteratur efter 1950, der er værd at læse?" af Nils Gunder Hansen (2022) samt tekstuddrag fra henholdsvis "Fru Marie Grubbe" af J. P. Jacobsen (1876), "Nordkraft" af Jakob Ejersbo (2002) og "det nemme og det ensomme" af Asta Olivia Nordenhof (2013).
Som udgangspunkt var ønsket at vedlægge disse tekster i undersøgelsen, så processen blev så virkelighedsnær for respondenterne som muligt og dermed ville højne projektets økologiske validitet. Med andre ord ville dette kunne styrke undersøgelsens gyldighed. Dog ville dette have medført brud på ophavsretten, hvorfor det ikke var ladsiggørligt. Dette fremgår af surveyen, hvor gymnasielærerne i stedet henvises til at tilgå teksterne via Prøvebanken, hvis de finder det relevant for deres besvarelse.

I surveyeksperimentet vil alle respondenter desuden blive præsenteret for den samme opgavebesvarelse fra en 3.g-elev. Dette sker med henblik på at kunne måle bedømmelsernes korrekthed, herunder niveauet af bias og støj, på tværs af grupperne.

Opgavebesvarelsen er en reel aflevering fra en 3.g-elev, som projektet har fået samtykke til at benytte i surveyeksperimentet. Besvarelsen kan ses i bilag 2. Gymnasieeleven har lavet besvarelsen som en afleveringsopgave og ikke i en egentlig eksamenssituation. Dog vurderes dette ikke at have nogen betydning for undersøgelsens mål, som i højere grad har fokus på gymnasielærernes bedømmelse. Det vil desuden sige, at opgavebesvarelsen er sket på baggrund af den egentlige opgaveformulering, som projektet ikke kan benytte i undersøgelsen af risiko for brud på ophavsret. Dog er omskrivningen af opgaveformuleringen sket med fokus på at beholde det samme indhold og naturligvis angive de samme eksterne tekster.

Gymnasieeleven, som har lavet besvarelsen, går på et af de 15 gymnasier, som afprøver en karakterfri gymnasieuddannelse. Derfor har besvarelsen som bedømmelse ikke modtaget en karakter fra 7-trins-skalaen, men "Godt ++" (bilag 2), hvilket ifølge eleven og dennes lærer vil svare til et stort 7-tal. Desuden har gymnasieeleven modtaget skriftlige nedslagspunkter i form af feedback gennem opgaven. Med udsagn som *"FY", "Flot", "Såh?"* og *"OK"* kan der med rimelighed argumenteres for, at eleven modtager feedback af minimal tilfredsstillende karakter. Afslutningsvist angiver gymnasielæreren feedback samt "karakter": *"Godt++. Velskrevet - spændende men genren?! Mere refleksion - mere jeg - mere undren"* (Bilag 2).

Denne opgavebesvarelse af 3.g-eleven er valgt netop på baggrund af denne bedømmelse. Projektgruppen finder denne feedback- og karaktergivning sigende for den problemstilling, som blev præsenteret i problemfeltet. Prøveordningen med karakterfri gymnasier kan siges at fordre endnu mere fyldestgørende feedback for at kompensere for den manglende karaktergivning, som ellers giver et systematisk indblik i bedømmelsen.

Desuden er denne besvarelse udvalgt, da den (hvis den var bedømt efter 7-trins-skalaen) er bedømt til et 7-tal. Som angivet i problemfeltet gives karakteren 7 for den gode præstation, som demonstrerer opfyldelse af fagets mål - men med en del mangler (UVM, 2023b). Projektet ønsker netop en god besvarelse, som også rummer fejl, da dette kan afhjælpe undersøgelsen af subjektivitet ved karaktergivning.

4.4.4 Brug af AI i surveyeksperiment

Til at give den udvalgte opgave feedback, har projektet valgt at anvende Google DeepMinds sprogmodel Gemini. Den opgraderede version af Gemini, Gemini Advanced, er baseret på den nye model Gemini 1.0 Ultra, som er den første chatbot-model der har kunne udkonkurrere

mennesker og eksperter i den såkaldte MMLU (Massive Multitask Language Understanding) test. Dette er en test, der undersøger evnen til at multitaske og løse problemer, som bruges til at teste diverse sprogmodeller. Denne test består af otte forskellige benchmarks, hvoraf Gemini rangerer højest på syv målt imod OpenAIs GPT-4, som er anvendt af ChatGPT (Google DeepMind, 2024). Projektgruppen vurderer derfor Gemini Advanced som den bedste nuværende løsning til at opfylde de krav, som en kunstig genereret karaktergivning skal udfylde.

Til denne sprogmodel har projektgruppen udarbejdet et prompt, der bedst muligt skal besidde de nødvendige data samt forståelse af danskfaget til at angive feedback til 3.g-elevens opgavebesvarelse. Dette prompt fremgår af bilag 3. Med dette prompt forsøger projektet at undgå for mange problematikker med generativ kunstig intelligens' faldgrube om den sorte boks. Dog må det også erkendes, at som beskrevet i problemfeltet, er der meget data i modellen, som der ikke kan redegøres for. Derfor vil det output, der anvendes i surveyeksperimentet, ikke kunne anses som fejlfrit. Det skjulte lag i Geminis neurale netværk er ikke muligt at se igennem, og det output, der anvendes, må derfor forstås som sit eget eksperiment. Geminis output er derfor skabt med henblik på at kunne udforme en sprogmodel, der kun kan anvende viden og data, som er relevant for at give karakter og agere som beslutningsstøtteværktøj for gymnasielærere, der underviser i dansk på A-niveau.

Som det fremgår af bilag 4, har Gemini Advanced ud fra projektet prompt, bedømt 3.g-elevens danskopgave til karakteren 7. Som det fremgår af forrige afsnit, stemmer dette overens med den karakter, som eleven rent faktisk fik af sin lærer. Karakteren 7 vil derfor fungere som den sande værdi for resten af projektets undersøgelse.

4.4.5 Surveyeksperimentets intervention

Efter en introduktion af projektets inddeling af respondenter samt en præsentation af undersøgelsens opgavebeskrivelse og -besvarelse, vil det følgende omhandle surveyeksperimentets intervention.

To af surveyeksperimentets grupper vil blive udsat for en intervention, og dermed fungere som eksperimentgrupper. Interventionen vil træde i kraft, efter de to eksperimentgrupper har læst 3.g-elevens opgavebesvarelse. Interventionen består som udgangspunkt af det samme: AI-genereret feedback samt en samlet karakter for opgavebesvarelsen. Det vil sige, at respondenterne i disse grupper bliver præsenteret for dette forud for egen feedback- samt karaktergivning. Kontrolgruppen skal angive karakter og feedback uden intervention, hvorfor

det forventes, at karaktergivningen her vil afspejle Kahneman et al.s teori om støj (2021) og CEPOS' analyse af karaktergivningen i gymnasiet (2021). Dette sker med en antagelse om, at de to eksperimentgruppers karaktergivning vil være mere korrekt end kontrolgruppens. Surveyeksperimentets intervention i form af Geminis feedback- og karaktergivning, som er vedlagt bilag 4, er desuden præsenteret nedenfor i figur 8:

Feedback samt karakter

Karakter: 7

Feedback:
Din opgave er velskrevet og reflekterende, og du har taget fat i de centrale aspekter af opgavebeskrivelsen. Du har en god forståelse af de forskellige kvaliteter og stilistiske egenskaber, der adskiller ældre og nyere litteratur, og du har inddraget relevante eksempler fra tekstuddragene "Fru Marie Grubbe" og "Nordkraft".

Styrker:
- Du har en klar og velformuleret introduktion, der præsenterer din centrale pointe.
- Du har en god struktur og overblik over opgaven, og du behandler alle de relevante delpunkter i opgavebeskrivelsen.
- Du har inddraget relevante citater fra tekstuddragene, og du bruger dem effektivt til at underbygge dine argumenter.
- Du har en god evne til at reflektere over dine egne tanker og oplevelser, og du deler dine personlige meninger på en velovervejet måde.
- Du har et godt sprog og en god formuleringsevne
- Du har en god evne til at skrive dialogisk og reflekterende. Du bruger et varieret sprog, og du stiller spørgsmål, der inviterer læseren til at reflektere over dine argumenter.

Forslag til forbedring:
- Du kunne have uddybet din analyse af "Fru Marie Grubbe" og "Nordkraft" yderligere. Du kunne f.eks. have inddraget flere detaljer fra teksterne, eller du kunne have analyseret de forskellige fortælleteknikker grundigt.
- Du kunne have konkluderet din opgave mere tydeligt. Du kunne f.eks. have opsummeret dine vigtigste pointer og have givet din endelige vurdering af, om nyere litteratur kan måles op mod ældre litteratur.

Din opgave er en god præstation, og du viser en god forståelse af emnet. Med en smule mere dybdegående analyse og en mere tydelig konklusion kunne din opgave have nået et endnu højere niveau.

Figur 8: Feedback samt karakter angivet af kunstig intelligens (bilag 4)

Eksperimentgruppernes intervention er udarbejdet på baggrund af Kahneman et al.s (2021) teori om støj i bedømmelser. Her argumenterer Kahneman et al. (2021) for, at det er muligt at påvirke bedømmeres niveau af bias ved at gøre brug af redskabet *debiasing*. Herunder vil interventionen kunne betragtes som *nudging*, der ved at gøre brug af *biasing information*, modificerer miljøet for bedømmelsen, influerer bedømmeren og reducerer eller tiltrækker bias

i en given retning. Denne *biasing information* kan desuden ses som det, man i fagsprog kalder en *second opinion*, som i dette tilfælde ikke er en kollegas, men genereret af generativ kunstig intelligens. Ved at præsentere respondenterne for den AI genererede *second opinion* vil de formodentlig blive påvirket og derfor læne sig op af denne.

Forskellen i de to gruppers interventioner består i præsentationen af ovenforstående beslutningsstøtteværktøj. Den ene gruppe bliver ikke oplyst om, at feedback samt karakteren er genereret af AI, mens den anden eksperimentgruppe bliver oplyst om beslutningsstøtteværktøjets ophav. Derfor formodes det, at eksperimentgruppen *Second Opinion Influence*, vil lade sig påvirke uden eventuelle fordomme om kunstig intelligens.

I *AI Bias* gruppen forventes det, at respondenterne vil være mere tilbageholdende med at stole på beslutningsværktøjet end SOI-gruppen. Dette antages på baggrund af Kahneman et al.s (2021) tese om fagprofessionelles objektive ignorance over for AI, da disse mennesker lever af at bruge deres dømmekraft og fagprofessionelle skøn.

Eksperimentgruppen SOIs introduktion til beslutningsstøtteværktøjet ses nedenfor:

Udklip 1: Eksperimentgruppe Second Opinion Influences introduktion til beslutningsstøtteværktøjet

Eksperimentgruppe AI bias' introduktion til interventionen fremgår nedenfor:

Udklip 2: Eksperimentgruppe AI bias' introduktion til beslutningsstøtteværktøjet

Disse antagelser er, på baggrund af teorien, forsøgt udmundet i undersøgelsens første hypotese:

H1a: Gymnasielærere, som har benyttet AI som beslutningsstøtteværktøj, vil i gennemsnit lave færre fejl i karaktergivningen i forhold til kontrolgruppen.

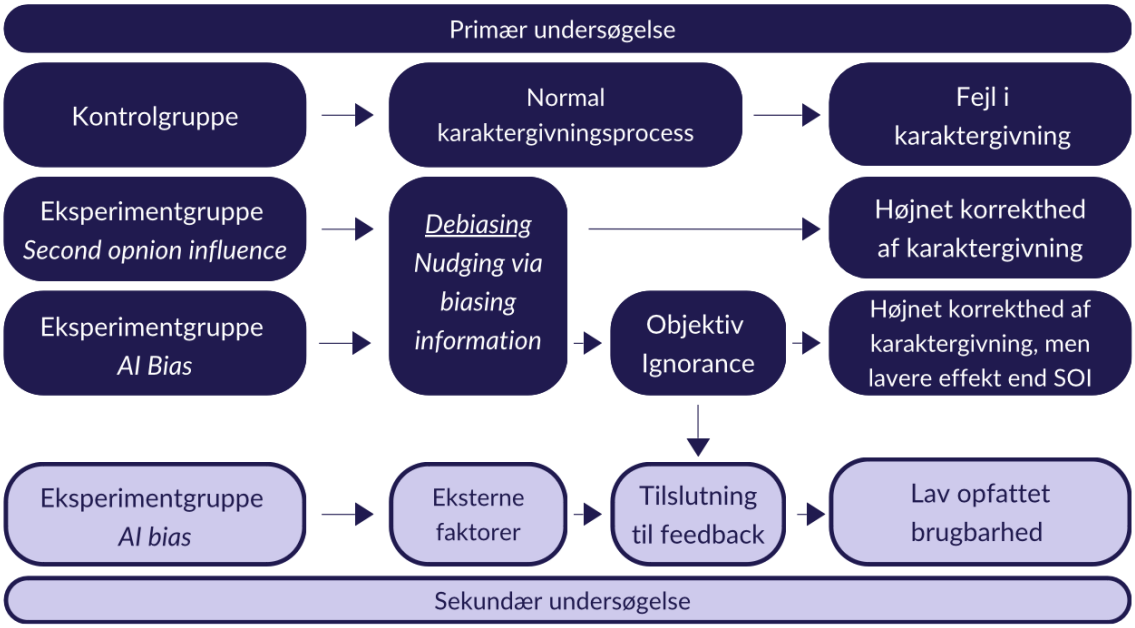
I forbindelse med surveyeksperimentet vil gymnasielærerne desuden svare på, om de er, har været eller aldrig har været censor. Dette spørgsmål vil kunne opdele gymnasielærerne i grupper, der har været udsat for *boosting*. På baggrund af tesen om at *boosting*, kvalificerer lærerne med censorerfaring til at foretage bedre og mere korrekte bedømmelser, vil det antages, at disse vil have en fordel, og dermed vil lave endnu færre fejl i bedømmelsen end dem, der ikke har censorerfaring. Dette leder frem til undersøgelsens anden hypotese:

H1b: Gymnasielærere, der er eller har været censorer og som anvender beslutningsstøtteværktøjet, vil have et lavere samlet fejlniveau i forhold til de lærere, der ikke har censorerfaring.

På baggrund af teorien om TAM (2000) og Kahneman et al.s (2021) tese om objektiv ignorance, vil der ydermere udformes en tredje hypotese. Antagelsen går på, at AI bias-gruppen vil have lavere tilslutning til den AI-genererede feedback og karakter end lærerne i SOI-gruppen, som ikke får at vide, at det er AI-genereret. Som nævnt tidligere vil fagprofessionelle værne så meget om deres fagprofessionelle skøn, at de vil afvise al teknologi, der udfordrer denne funktion. Dette leder til følgende hypotese:

H2: Eksperimentgruppen AI bias vil være mere tilbøjelig til at være uenige med feedback- og karaktergivning genereret af AI i forhold til eksperimentgruppen Second Opinion Influence.

Forventningerne til surveyeksperimentets intervention er visualiseret gennem nedenstående figur 9:



Figur 9: Visualisering af projektets undersøgelser

4.4.6 Surveyeksperimentets stikprøve

I dette afsnit vil der redegøres for projektets surveyundersøgelse. Herunder antallet af distributioner, det endelige antal respondenter samt den demo- og geografiske repræsentativitet. I forbindelse med distribution af projektets surveyeksperiment har målet været at generere den størst mulige stikprøve i populationen. Dette ønske er naturligt i forbindelse med ønsket om høj repræsentativitet. Distributionen har som udgangspunkt haft til mål at generere 15 respondenter i hver af de tre grupper i surveyeksperimentet, da dette kræves for at sikre på bedst mulig vis, at stikprøven fordeles tæt på en normalfordeling ved hypotesetest (Agresti, 2018: 137).

Projektets surveyeksperiment endte med i alt 461 deltagere, hvoraf 119 af disse gennemførte hele surveyen. Som det blev præsenteret i 4.4.2 og tabel 4, er det lykkedes at få mere end 15 respondenter i hver gruppe.

Projektets surveyeksperiment indeholder desuden kontrolvariable, som udgøres af køn, region, års erfaring og hvorvidt gymnasielærerne er censorer eller ej. Disse har primært til formål at undersøge repræsentativiteten af surveyeksperimentets stikprøve, men udvalgte variable vil

også indgå i dele af analysen. Af hensyn til projektets økologiske validitet findes det relevant at udfolde undersøgelsens stikprøve.

En stikprøve er et udsnit af den samlede population, som tager del i genstandsfeltet (Clement & Ingemann, 2019: 97-99). En stikprøve er fordelagtig, da det oftest vil være praktisk umuligt at undersøge hele populationen. Det er afgørende for projektets eksterne validitet, at stikprøven er repræsentativ og dermed kan findes et andet sted i populationen. Denne test af stikprøvens repræsentativitet vil ske gennem en stratificeret stikprøve (Clement & Ingemann, 2019: 98-99).

Stikprøven består som nævnt af 119 gymnasielærere i dansk på A-niveau. Surveyeksperimentets kontrolvariable benyttes til at undersøge repræsentativiteten af denne stikprøve i sammenligning med populationen, som består af alle danske gymnasielærere. Disse data findes umiddelbart ikke tilgængelige, hvorfor projektgruppen har kontaktet Gymnasielærernes Landsforening samt Børne- og Undervisningsministeriet for at modtage disse. Disse data ses i nedenstående tabel 5:

Kønsfordeling		
	Surveyeksperimentets stikprøve	Fordelingen blandt medlemmer af GL
Mand	34,5%	28%
Kvinde	65,5%	72%
Geografisk fordeling		
	Surveyeksperimentets stikprøve	Fordelingen blandt medlemmer af GL
Nordjylland	16%	9,5%
Midtjylland	14%	24,5%
Syddanmark	23%	19%
Sjælland	19%	11%
Hovedstaden	28%	34%
Ukendt	-	2%
Censorfordeling		
	Surveyeksperimentets stikprøve	Fordelingen blandt medlemmer af GL
Er eller har været censor	67%	30,5% *
Har aldrig været censor	33%	Ukendt

**Andelen af dansklærer, som er censorer på nuværende tidspunkt (UVM)*

Tabel 5: Surveyeksperimentets stikprøve og populationen

Gennem korrespondance med GL, har projektgruppen modtaget data over samtlige gymnasielærer, som underviser i dansk på A-niveau enten som hovedfag eller bifag.
Af ovenstående tabel ses det, at stikprøvens kønsfordeling er relativt tæt på den, der er i populationen. Dog har denne stikprøve 6,5 procentpoint færre kvindelige respondenter, end det er tilfældet i populationen. På baggrund af dette parameter findes stikprøven repræsentativ for populationen.

Den geografiske fordeling af respondenter findes ligeledes relativt repræsentativ for populationen. Andelen af respondenter fra Region Midtjylland afviger mest fra populationen. Her er stikprøvens størrelse 10,5 procentpoint mindre end populationen. Dog findes denne afvigelse ikke af mærkbar karakter, hvorfor stikprøven findes repræsentativ for populationen.
Afslutningsvist ses fordelingen af censorer. Disse data har ikke været lige så nemt at tilvejebringe som køn og region. Projektgruppen har gennem Børne- og Undervisningsministeriet modtaget data for, hvor mange nuværende gymnasielærere, som underviser i dansk på A-niveau, der *er* censorer. Det har ikke været muligt at få data for, hvor mange dansklærere i gymnasiet, som *har været* censorer. Det vil sige, at andelen på 30,5 procent udgøres af nuværende gymnasielærere i dansk, som er censorer. Denne andel er beregnet ved at dividere det samlede antal censorer, som projektgruppen har modtaget fra GL med data om nuværende censorer, som er modtaget af Børne- og Undervisningsministeriet. Derfor findes det ikke relevant at vurdere stikprøvens repræsentativitet på baggrund af dette parameter, da dataindsamlingen er mangelfuld.

Samlet set vurderes stikprøven som repræsentativ for populationen, hvilket primært vurderes på baggrund af respondenternes fordeling i køn og region.

4.5 Operationalisering

For at gøre undersøgelsens resultater målbare i samspil med de udvalgte teorier samt forskningsdesignets udfoldelse, vil der i dette afsnit redegøres for surveyeksperimentets variable samt den teoretiske forståelse af, hvordan en bedømmer begår fejl. Surveyen har i alt 119 respondenter og elleve variable. Alle respondenterne er blevet fordelt i tre grupper hvor to af dem udsættes for en intervention i form af projektets beslutningsstøtteværktøj. De forskellige variable vil blive undersøgt nærmere ved hjælp af statistikprogrammet STATA. Denne fremgangsmåde vil bruges til at besvare de opstillede hypoteser samt problemformuleringen.

4.5.1 Præsentation af surveyeksperimentets variable

I dette afsnit vil de variable, som benyttes i projektets analyse, præsenteres. Alle variable er fremkommet gennem spørgsmål i projektets survey (bilag 1). Præsentationen vil indebære en karakterisering af variablens type og skalering samt eventuelle rekodninger af svarkategorier. Formålet med dette er at være transparente omkring fordelingen af respondenter gennem surveyeksperimentet. Dertil ønsker projektgruppen ikke at manipulere data gennem diverse rekodninger, hvorfor disse vil fremgå eksplicit i dette afsnit.

Den afhængige variabel gennem hele analysen er respondenternes angivne `karakter` for 3.g-elevens opgavebesvarelse. Respondenterne i surveyeksperimentet bliver afslutningsvist bedt om at angive en samlet karakter for opgavebesvarelsen "Reflekterende artikel: Nyere og ældre litteratur". Svarkategorierne her er 7-trins-skalaens syv karakterniveauer. Denne variabel er intervalskaleret, da kategorierne er naturligt rangordnet efter 7-trins-skalaen. Frekvensen for besvarelser i denne variabel fremgår af følgende udskrift fra STATA, som ses nedenfor i tabel 6.

Afhængig variabel: Karakter		
	Frekvens	Andel
12	3	2,52%
10	24	20,17%
7	54	45,38%
4	34	28,57%
02	3	2,52%
00	1	0,84%
-3	0	0%

Tabel 6: Frekvens af karakterer i surveyeksperimentet

I projektets anden delanalyse vil der være to primære uafhængige variable, som er de to eksperimentgruppers holdning til den angivne feedback, der hedder henholdsvis `bf_SOI` og `bf_AIbias`. I surveyeksperimentet bedes respondenterne i SOI-eksperimentgruppen besvare *"I hvilken grad kan du tilslutte dig den givne feedback samt karakterbedømmelse?"*, mens AI

bias gruppen bedes besvare *"I hvilken grad kan du tilslutte dig den givne feedback samt karakterbedømmelse, som er genereret af kunstig intelligens?"*. Hertil er svarkategorierne forudbestemte som "Meget lav", "Lav", "Hverken/eller", "Høj" og "Meget høj". Disse variable er begge ordinalt skalerede, da de er naturligt rangordnede, men der kan ikke måles en afstand mellem dem.

Nedenfor ses tabel 7 som en frekvenstabel over fordelingen af respondenter i *Second Opnion Influence* gruppens tilslutning til den angivne feedback, som er præsenteret som et beslutningsstøtteværktøj.

Primær uafhængig variabel: SOI tilslutning til feedback		
	Frekvens	Andel
Meget høj	0	0%
Høj	13	35,14%
Hverken/eller	9	24,32%
Lav	11	29,73%
Meget lav	4	10,81%
Total	**37**	**100%**

Tabel 7: Eksperimentgruppe Second Opinion Influences tilslutning til den angivne feedback, som er genereret af AI

For at kunne benytte denne variabel i en hypotesetest i analysen, rekodes denne til en dummy variabel, der kaldes `SOI_dum` i STATA. Derfor kodes kategorierne "Meget lav" og "Lav" sammen og "Meget høj" og "Høj". "Hverken/eller" kodes missing, da denne svarkategori ikke findes relevant for analysens resultater. Som det fremgår nedenfor i tabel 8, vil variablen efter denne kodning se således ud:

Primær uafhængig variabel: Dummy SOI tilslutning til feedback		
	Frekvens	Andel
Meget høj / Høj	13	46,5%
Meget lav / Lav	15	53,5%
Total	**28**	**100%**

Tabel 8: Eksperimentgruppe Second Opinion Influences (dummy) tilslutning til den angivne feedback, som er genereret af AI

Nedenfor ses tabel 9, som er en frekvensstabel for fordelingen af respondenter i AI bias gruppens tilslutning til den angivne feedback, som de ved, er genereret af AI.

Primær uafhængig variabel: AI bias tilslutning til feedback		
	Frekvens	Andel
Meget høj	1	2,04%
Høj	13	26,53%
Hverken/eller	13	26,53%
Lav	17	34,69%
Meget lav	5	10,2%
Total	**49**	**100%**

Tabel 9: Eksperimentgruppe AI bias' tilslutning til den angivne feedback, som er genereret af AI

Denne ordinalt skalerede variabel kodes på samme måde som den foregående. Som det fremgår af tabel 10, vil dummyvariablen `AIbias_dum` efter rekodningen se således ud:

Primær uafhængig variabel: Dummy AI bias tilslutning til feedback		
	Frekvens	Andel
Meget høj / Høj	14	39%
Meget lav / Lav	22	61%
Total	**36**	**100%**

Tabel 10: Eksperimentgruppe AI bias' (dummy) tilslutning til den angivne feedback, som er genereret af AI

Surveyeksperimentet har desuden fire variable, som vil indgå som kontrolvariable til både undersøgelsens stikprøve samt i analysen.

Den første uafhængige variabel er `køn`, som er en dikotom variabel, hvilket betyder, at der er to kategorier. Fordelingen af respondenterne mellem disse to kategorier ses i nedenstående frekvenstabel i tabel 11:

Uafhængig variabel: Køn		
	Frekvens	Andel
Mand	41	34,45%
Kvinde	78	65,55%
Total	**119**	**100%**

Tabel 11: Køn

Variablen `region` er nominelt skaleret, hvor Danmarks fem regioner udgør de forskellige kategorier, som ikke kan naturligt rangordnes. Frekvenstabellen for denne variabel ses nedenfor i tabel 12.

Uafhængig variabel: Region		
	Frekvens	Andel
Nordjylland	19	15,97%
Midtjylland	17	14,29%
Syddanmark	27	22,69%
Sjælland	23	19,33%
Hovedstaden	33	27,73%
Total	**119**	**100%**

Tabel 12: Region

Den uafhængige variabel `erfaring` er intervalskaleret, da kategorierne kan rangordnes, og det er desuden muligt at beregne den præcise afstand mellem disse. Frekvenstabellen for denne variabel ses nedenfor i tabel 13:

Uafhængig variabel: Erfaring		
	Frekvens	Andel
0-5 år	12	10,08%
6-10 år	18	15,13%
11-15 år	28	23,53%
16-20 år	26	21,85%
+20 år	35	29,41%
Total	**119**	**100%**

Tabel 13: Erfaring

Den uafhængige variabel `censor`, omhandler hvorvidt respondenterne er censor, har været censor eller aldrig har været censor. Denne variabel er nominelt skaleret, og frekvenstabellen for fordelingen ses nedenfor i tabel 14:

Uafhængig variabel: Censor		
	Frekvens	Andel
Er censor	46	38,6%
Har været censor	34	28,57%
Har aldrig været censor	39	32,77%
Total	**119**	**100%**

Tabel 14: Censor

For at kunne benytte denne variabel i forbindelse med hypotesetests i analysen, kodes den om til en dummy. Her kodes de to grupper "Er censor" og "Har været censor" sammen til én gruppe, da disse efter teori fra Kahneman et al. (2021) alle har været udsat for *boosting*, mens "Har aldrig været censor" ikke har været udsat for *boosting*.

Som det fremgår af tabel 15, ser dummyvariablen efter rekodning, således ud:

Uafhængig variabel: Dummy censor		
	Frekvens	Andel
Har været/er censor	80	67,23%
Har aldrig været censor	39	32,77%
Total	**119**	**100%**

Tabel 15: Censor (dummy)

Afslutningsvist benytter projektet gennem analysen den uafhængige variabel `s_5`, som består af gymnasielærernes kvalitative feedback på 3.g-elevens opgavebesvarelse. Udsagnene fra denne variabel er i STATA opdelt i de tre grupper *Second Opinion Influence*, AI bias og kontrolgruppen, og kan ses i bilag 5.

4.5.2 Måling af fejl i bedømmelser

Når det kommer til målingen af de fejl, som potentielt forekommer i bedømmelsespraksis, bliver det mere kompliceret. Her skelner Kahneman et al. (2021) imellem to forskellige typer af bedømmelsespraksisser. Forudsigende bedømmelser og evaluerende bedømmelser.

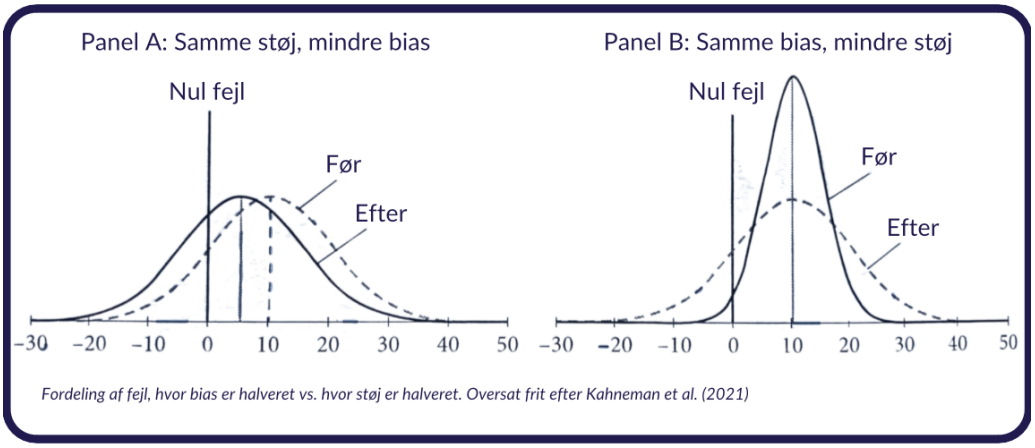
Forudsigende bedømmelser er vurderinger, hvor der kan findes en "sand" værdi. Eksempelvis når et ishus skal vurdere, hvor meget is, der bliver solgt i løbet af weekenden. Når weekenden er gået, kan ishuset se, hvor langt fra målet de vurderede. Lavede de for meget is, eller lavede de for lidt is? Herefter kan ishuset justere produktionen til næste weekend, så de forhåbentligt rammer målet mere akkurat og præcist.

Den evaluerende bedømmelse er sværere, fordi der ikke er nogen decideret "sand" værdi. Hvad er værdien af maleriet? Hvilken score skal kunstskøjteløberen have? Hvilken karakter skal eleven have for sin opgave? Alle disse bedømmelser beror på smag, hvor der sjældent er forventet enighed (Kahneman et al., 2021: 362). Hvis der tilnærmelsesvis kan etableres en "sand" værdi, så kan fejlene ikke vægtes lige meget i begge retninger. Dette skal forstås på den måde, at fejlen ved at give en for høj karakter er mindre betydningsfuld (og dermed mindre vægt) end fejlen ved at give en for lav karakter (Kahneman et al., 2021: 67).

Dog vil der i eksperimentet blive gjort brug af kunstig intelligens, der har til formål at give et objektivt svar på, hvilken karakter, som er passende for danskopgaven, der bruges. Som nævnt i afsnit 3.1.10, præsenterer Kahneman et al. (2021) selv, at AI og algoritmer er objektive og støjløse, hvorfor denne karakter må antages at være den mest sande værdi for danskopgaven (Kahneman et al., 2021: 124; 334). Derfor vil der i det følgende afsnit argumenteres for, at *fejlligningen* som Kahneman et al. (2021) præsenterer til måling af det samlede fejlniveau ved forudsigende bedømmelser, kan benyttes. Dette er på trods af, at bedømmelsespraksissen for gymnasielærere i dansk er en evaluerende bedømmelse. Dette betyder endvidere, at fejl på begge sider af den sande værdi vil blive vægtet lige højt, selvom det i projektets genstandsfelt ikke er sådan i realiteten. Der er relativt samme afstand fra 00 til 7, som afstanden fra 7 til 12, men fejlen har større konsekvens for eleven, hvis der gives en lavere karakter i forhold til en højere karakter. Det anerkendes, men ses bort fra, at der er forskel i fejlstørrelse afhængigt af, hvilken side af den sande værdi, fejlen befinder sig på (Kahneman et al., 2021: 67).

Målet ved karaktergivningsprocessen er at opnå det mest objektive resultat, hvor hverken bias eller støj fører til skævvridninger i karaktergivningen. Med viden om bias (konstante, forudsigelige fejl) og støj (uforudsigelige fejl), kan det bestemmes, hvor meget henholdsvis bias og støj bidrager til de fejl, der bliver foretaget. Eksempelvis kan en gymnasielærer konsekvent undervurdere kvaliteten af danskopgaver en lille smule, men en gang imellem overvurderer læreren kvaliteten ekstremt meget. Fejlene, hvad enten der er tale om bias eller støj, udligner ikke hinanden - de hober sig bare op (Kahneman et al., 2021: 55). Det er vigtigt at kunne skelne bias og støj fra hinanden, da de er uafhængige af hinanden, hvad angår det samlede fejlniveau. Dog giver en reduktion af enten bias eller støj den samme effekt på det samlede fejlniveau.

Kahneman et al. (2021) benytter Gaussfordelingen (normalfordelingen) til at visualisere fejl i bedømmelsespraksis (Kahneman et al., 2021: 56). Denne visualisering i figur 10 kan hjælpe til at bestemme, hvorvidt der reduceres bias, støj eller begge dele.



Figur 10: Normalfordeling af fejl, hvor bias er halveret versus hvor støj er halveret - oversat frit efter Kahneman et al. (2021)

I Gaussfordelingen er gennemsnittet af bedømmelserne i spidsen af normalfordelingen. Den gennemsnitlige bedømmelse kan subtraheres fra den sande værdi for at beregne, hvor langt den gennemsnitlige lærer er fra målet. Denne forskel er et udtryk for den første komponent i det samlede fejlniveau; bias. Hvis forskellen mellem gennemsnittet og den sande værdi er blevet mindre, så er der i gennemsnit mindre bias i bedømmelserne (Kahneman et al., 2021: 58). Derved kan det bestemmes om lærerne i gennemsnit bedømmer danskopgaver hårdt eller generøst på baggrund af hvilken side, gennemsnittet befinder sig på i forhold til den sande værdi, som er bestemt af kunstig intelligens (Kahneman et al., 2021: 59).

Den anden komponent i det samlede fejlniveau er støj. Støj er ifølge Kahneman et al. (2021) lig med standardafvigelsen i Gaussfordelingen, da den er et udtryk for, hvor meget den gennemsnitlige lærers bedømmelse afviger fra gennemsnittet. Denne kan også kaldes for spredningen i karaktererne i stikprøven (Kahneman et al., 2021: 57). Hvis spredningen er stor, så er lærerne meget støjende og upræcise i deres bedømmelse.

Kahneman et al. (2021) bruger Carl Friedrich Gauss' metode til at bestemme det samlede fejlniveau, hvor både støj og bias kombineres til én samlet måleenhed. Denne måleenhed kaldes *Mean Squared Error* (MSE), og er forskellen mellem den sande værdi og bedømmelsen opløftet i anden potens (Kahneman et al., 2021: 59). Grunden til, at værdien opløftes i anden potens, er for at vægte størrelsen af fejlen. Hvis den sande værdi er et 7-tal, så er et 4-tal en mindre fejl end 02. Denne vægtning giver store fejl større betydning end små fejl (Kahneman et al., 2021: 61). Denne metode er den foretrukne til at bestemme det samlede fejlniveau, hvor

præcision er målet. Jo tættere MSE er på 0, desto færre fejl er der i bedømmelsen. Metoden kommer af *fejlligningen*, der siger:

$$Fejl\ i\ en\ enkelt\ måling\ =\ Bias\ +\ Støj$$

I dette tilfælde, hvor flere bedømmelser af samme sag er givet, vil *fejlligningen* udvides til følgende:

$$Samlede\ fejlniveau\ = Bias^2\ + Støj^2$$ (Kahneman et al., 2021: 62).

I denne ligning er bias angivet som forskellen mellem den gennemsnitlige bedømmelse og den sande værdi, og støjen er afgivet som standardafvigelsen.

Denne ligning beror på Pythagoras sætning, der siger, at summen af to kvadrater, der udgør en retvinklet trekants korte sider, udgør kvadratet på den længste side. På den måde vægtes støj og bias ligeligt, når der er tale om det samlede fejlniveau. Hvis den ene eller den anden komponent reduceres, vil det have den samme effekt på det samlede fejlniveau (Kahneman et al., 2021: 63). Dette ses illustreret i figur 11, der viser, at MSE er afhængig af størrelserne på de to komponenter på den retvinklede trekants korte sider, som er henholdsvis bias i anden potens og støj i anden potens. Bliver enten mængden af bias eller støj reduceret, vil det samlede fejlniveau (MSE) ligeledes blive reduceret.

Figur 11: To dekompositioner af DSE - fortolket frit efter Kahneman et al. (2021)

Er det så lige meget, om det er støj eller bias, der reduceres? Ønskescenariet er selvfølgelig at reducere begge dele. Der er imidlertid stor forskel på udfaldet af bedømmelserne afhængigt af, om der interveneres for at reducere bias eller støj.

Hvis støjen reduceres, men bedømmernes bias forbliver det samme, så bliver bedømmelserne mere præcise. Dog er der nu en større procentdel, hvis bedømmelse afviger fra målet. Dette er næsten værre end før interventionen, da det bliver en "mere præcist forkert bedømmelse". Dette bevirker til gengæld, at den eksisterende bias bliver endnu mere tydelig og forhåbentligt adresseret som noget, der må ændres efterfølgende (Kahneman et al., 2021: 64-66).

Hvis der sker en reducering i bias, vil spredningen i bedømmelserne være den samme, men flere rammer tættere på målet. Dog forekommer der stadig store fejl på begge sider af den sande værdi (Kahneman et al., 2021: 64).

I henhold til MSE vil det svare til den samme reducering i fejl uanset hvilken komponent, der reduceres. Derfor må det efterstræbes at minimere de store fejl, da store fejl vægter højere grundet fejlligningen og dens vægtning af fejl i anden potens (Kahneman et al., 2021: 65).

4.5.3 Fejlkilder

Projektgruppen blev efter surveyeksperimentets distribution gjort opmærksom på flere mulige fejlkilder. For at projektets undersøgelse af de opsatte hypoteser kan siges at være gyldig, er det relevant at gøre opmærksom på disse fejlkilder, da de kan have indflydelse på den målte effekt og dermed projektets interne validitet.

4.5.3.1 Udformningen af surveyeksperimentet

Indsamlingen af relevante gymnasielæreres mails er som nævnt sket manuelt gennem alle Danmarks gymnasiers liste over ansatte på de respektive hjemmesider. Denne del af distributionen fordrer, at gymnasiernes hjemmesider er opdaterede i forhold til, om respondenterne opfylder kriterierne for deltagelse.

Projektgruppen har efter distribution modtaget adskillige mails fra respondenter, som ikke længere er undervisere i dansk, hvorfor de ikke opfylder kriterier for deltagelse. For at afværge denne fejlkilde er de frafaldet som respondenter i undersøgelsen i SurveyXact.

Projektgruppen blev desuden gjort opmærksom på forskelle i opgavetyper på de forskellige gymnasiale uddannelser, som undersøgelsen er udsendt til. Ved surveyeksperimentets

distribution blev både gymnasielærere på stx, hhx, htx og hf valgt, da disse alle underviste i dansk på A-niveau. Dog er det siden blevet klargjort, at uddannelsen hf ikke gør brug af opgavetypen "reflekterende artikel", hvorfor lærerne med tilknytning til denne uddannelse har svære forudsætninger for at besvare undersøgelsen. Derfor er disse desuden frafaldet som respondenter af undersøgelsen i SurveyXact.

Projektets opdeling af respondenter i henholdsvis eksperimentgrupper og kontrolgrupper sker på baggrund af deres fødselsdagsmåned. Det er muligt, at denne metode til inddeling af respondenterne kan være med til at skabe sampling bias (Aarhus Universitet, 2024). Det betyder, at denne tilfældige inddeling kan give tilfældige skævheder i de tre surveygrupper i kontrolvariablene, så grupperne får en skæv demografisk repræsentativitet. Desuden giver denne tilfældige inddeling heller ikke kontrol over, hvor mange respondenter, som ender i hver gruppe.
Den potentielle fejlkilde ved sampling bias ses undersøgt gennem kontrol af surveyeksperimentets stikprøve i afsnit 4.4.6.

4.5.3.2 Situationsstøj

Når der foretages eksperimenter, er det vigtigt at holde så mange indvirkende faktorer konstante som muligt, så det sikres, at effekten skyldes interventionen - og kun den. Dette er dog ualmindeligt svært, når der som i dette projekt undersøges psykologiske processer i hjernen, som resulterer i udfald af bedømmelsesprocessen. Således er det også umuligt at eliminere situationsstøjen, som unægteligt altid vil foregå i beslutningssitutioner. Selv med denne erkendelse, er det næsten umuligt at enten begrænse den eller påvirke den i en given retning. Individer vil altid være påvirket af forskellige sindstilstande i forskellige situationer, som skaber den forbigående, tilfældige støj, der ikke kan kontrolleres (Kahneman et al., 2021: 93). Dog er det tidligere nævnt i afsnit 3.1.5, at den stabile mønsterstøj er dominerende i forhold til situationsstøjen, som dermed kun udgør en lille del af den samlede mønsterstøj (Kahneman et al., 2021: 213). Derfor må denne utilsigtede påvirkning regnes som værende ubetydelig, men dog til stede.

4.5.3.3 Satisficing

Ved undersøgelser foretaget gennem surveys er det gældende, at respondenter ikke altid er motiverede eller i stand til at besvare de stillede spørgsmål. Begrebet satisficing dækker over

denne kognitive tendens, hvor en respondent stiller sig tilfreds med den første løsning, der løser problemet eller besvarer spørgsmålet - og ikke nødvendigvis den mest optimale eller fyldestgørende besvarelse (Clement, 2017: 104).

Sandsynligheden for *satisficing* afhænger primært af tre faktorer. Først og fremmest kan respondentens mangle de nødvendige evner til at kunne besvare et spørgsmål. Dernæst afhænger respondentens motivation til at besvare spørgsmålet også af graden af *satisficing*. Afslutningsvist nedsættes respondentens evner samt motivation til besvarelse af sværhedsgraden på spørgsmålsbatteriet (Clement, 2017: 107-108).

Før distribueringen af projektets surveyeksperiment til populationen af dansklærere i gymnasiet blev dette gennemset og afprøvet af en forhenværende gymnasielærer i dansk. Foruden at øge projektets økologiske validitet, var formålet med dette også at mindske risikoen for *satisficing*. Projektgruppen modtog herigennem konstruktiv kritik af flere dele af surveyeksperimentet om mangelfulde anvisninger eller dele af processen, som kunne optimeres således, at oplevelsen for respondenterne ville være så overskuelig og virkelighedsnær som muligt.

Her fik projektgruppen bekræftet, at omfanget af den opgave, som surveyeksperimentet stiller respondenterne, i nogen grad øger risikoen for *satisficing*. Den forhenværende gymnasielærer vedkendte selv at være sprunget lettere over dele af retteprocessen, end vedkommende ville gøre normalvis. Dette er desuden bekræftet gennem korrespondance med flere af gymnasielærerne, som har fremhævet, at distribueringen af surveyeksperimentet er sket i en særdeles travl periode for dem.

Anden delanalyses to primære uafhængige variable omhandler graden af tilslutning til den angivne feedback. Her er en af svarmulighederne "Hverken/eller". Projektgruppen anerkender, at denne form for svarkategori øger risikoen for *satisficing*. Det kræver mere af respondenten at tage egentlig stilling til graden af tilslutning, fremfor blot at besvare "Hverken/eller". Dog findes det heller ikke fordelagtigt at udelade denne svarkategori. I så fald risikeres det, at respondenter, som ikke er i stand til at tage stilling til tilslutningen, må se sig nødsaget til, at vælge imellem lav eller høj grad alligevel. (Clement, 2017: 106-107). Derfor er denne svarkategori en del af spørgsmålsbatteriet, men for at mindske graden af *satisficing* i projektets analyse er respondenterne, som har besvaret disse spørgsmål med "Hverken/eller" kodet missing.

4.5.3.4 Social desirability

Begrebet *social desirability* dækker over den tendens, hvor respondenters svar styres af sociale normer og generelt socialt accepteret adfærd (Clement, 2017: 109-111). Det betyder, at en respondent kan være tilbøjelig til at lade sin besvarelse styre af, hvad vedkommende mener vil være det rigtige i andre individers øjne.
Risikoen for *social desirability* imødekommes af, at det flere gange i processen for distribution understreges tydeligt, at respondenternes besvarelser forekommer fuldkomne anonyme. På den måde er det ikke muligt at kæde respondenternes holdninger og svar sammen med deres person.

Projektgruppen er opmærksom på risikoen for *social desirability* i forbindelse med eksperimentgruppen AI bias. Kunstig intelligens er hurtigt blevet en del af mange individers liv - herunder også skoleelever. Som skitseret i problemfeltet er flere undervisere negativt stemt overfor brugen af diverse sprogmodeller, da det i stigende grad benyttes til snyd af elever. Dermed kan der argumenteres for, at det imellem lærere er mest acceptabelt ikke at kunne tilslutte sig projektets beslutningsstøtteværktøj, som er genereret af kunstig intelligens.

Projektgruppen anerkender dette, da *social desirability* kan ses som en forudsætning for udformningen af H2 set i lyset af TAM. Teorien bag TAM præsenterer blandt andet den subjektive norm, som i høj grad baserer sig på de samme tendenser som *social desirability*.

4.6 Analysestrategi

Efter foregående præsentation af surveyeksperimentets variable, fejlkilder samt måling af fejlniveau (Kahneman et al., 2021), vil der i det følgende blive redegjort for projektets statistiske analyser samt brugen af kvalitative bidrag.
Alle statistiske analyser vil finde sted i analyseprogrammet STATA. Kommandoer kan aflæses i projektets do-file, som fremgår af bilag 6.

Fælles for alle dele af den statistiske analyse i STATA er, at disse vil blive foretaget indenfor et 90 procent konfidensinterval. Et konfidensinterval for en sammenhæng angiver et interval, som det antages, at stikprøven falder indenfor. Dette interval skal ideelt være tæt på værdien 1, hvorfor det ofte ses, at det anføres som 0,95 eller 0,99 (Agresti, 2018: 115-117). Dog vurderes stikprøvens størrelse at være relativt lille inden for en statistisk analytisk kontekst. For i højere grad at kunne opnå statistisk signifikante resultater, anføres da et sikkerhedsinterval på 0,90. Dette gøres ved betingelsen `level(90)` som afslutning på alle hypotesetests.

Ved et 90 procent konfidensinterval angives det ved 90 procents sikkerhed, at en given parameter ligger indenfor det angivne interval.

4.6.1 Missingfilter

For at øge mållingsværdien af projektets undersøgelse og resultater udarbejdes et missingfilter i STATA. Dette missingfilter har til formål at udelukke de respondenter, som ikke har besvaret surveyens sidste spørgsmål, hvor der skal angives en samlet karakter for 3.g-elevens opgavebesvarelse. Dette sker, da det ønskes, at projektets undersøgelse udelukkende beror på respondenter, som har besvaret hele surveyen. Disse respondenter kodes ud gennem kommandoen `drop`, som betinges af, at variablen `karakter` ikke er besvaret.
Efter genereringen af dette missingfilter er der 119 respondenter i undersøgelsen, hvilket også er tilfældet for frekvenstabellerne i præsentationen af variable i afsnit 4.5.1.

4.6.2 Undersøgelse af korrektheden af karaktergivning i surveyundersøgelsen

Analysen indledes ved at undersøge, hvordan respondenterne i projektets surveyeksperiment har bedømt 3.g-elevens opgavebesvarelse. På den måde klarlægges fra start, hvorvidt der forefindes fejl i surveyeksperimentets respondenters karaktergivning. Dette sker gennem brug af kommandoen `tab` i STATA efterfulgt af den afhængige variabel `karakter`. Den afhængige variabel `føds` tilføjes for at se fordelingen i karaktergivningen mellem de tre grupper.

Disse data behandles i Excel for at visualisere fordelingen gennem et søjlediagram og i Geogebra for at visualisere fordelingen gennem tre Gaussfordelinger. Grafen med Gaussfordelingerne har ydermere fået tilføjet fire vertikale linjer, der er sat ved henholdsvis den sande karakter (karakteren 7) og de tre gruppers respektive karaktergennemsnit. Denne visualisering gør det nemmere at se forskellene mellem grupperne og dermed interventionens effekt.

4.6.3 Kvalitative forklaringer

Resultaterne fra foregående statistiske databehandling vil elaboreres yderligere gennem inddragelse af kvalitative bidrag. Som nævnt i afsnit 4.3, så har den metodiske tilgang i dette projekt henblik på at opnå mixed methods-forskning. Gennem projektets surveyeksperiment

har projektgruppen modtaget feedback samt en samlet karakter til 3.g-elevens opgavebesvarelse fra alle respondenter. Disse udsagn stammer fra projektets surveyeksperiment, hvor variablen `s_5` indeholder alt den kvalitative feedback, som respondenterne har angivet til 3.g-elevens opgavebesvarelse. Al feedback opdeles i de to eksperimentgrupper og kontrolgruppen (bilag 5).

En kvalitativ analyse af gymnasielærernes feedback på 3.g-elevens danskopgave vil give et dybere indblik i, hvad der ligger til grund for de forskellige bedømmelser. Dette er med henblik på at bevise, at der opstår mønsterstøj i forbindelse med karaktergivningsprocessen. Her vil der blive lagt vægt på tekstpassager, der vidner om niveaustøj eller stabil mønsterstøj. Situationsstøjen er umiddelbart ikke mulig at aflæse i det angivne feedback, da det ikke direkte fremgår, hvilket humør gymnasielærerne er i, når de bedømmer opgavebesvarelsen.

4.6.4 Undersøgelse af beslutningsstøtteværktøjets effekt på karaktergivning

Til at besvare projektets primære hypotese (H1a) vil der foretages to-sidet hypotesetests af to omgange. Dette gøres ved kommandoen `ttest` i STATA. Formålet med hypotesetests er at teste for forskelle mellem to grupper for på den måde at undersøge problemstillingens sammenhæng. Der kræves minimum 15 respondenter i hver gruppe for at sikre, at stikprøven er tilnærmelsesvis normalfordelt (Agresti, 2018: 137).

En hypotesetest har til formål at be- eller afkræfte en sammenhæng, hvorfor der for hver hypotese implicit foreligger en nulhypotese. Denne antager, at der ikke er forskel mellem de to grupper, som testes mellem (Agresti, 2018: 152-153). Projektet opstiller ikke nulhypoteser gennem analysen, men disse vil som nævnt ligge implicit til hver af de tre opstillede hypoteser.

Der foretages først en hypotesetest for forskellen mellem karaktergivning for kontrolgruppen og eksperimentgruppen *Second Opinion Influence*. Dernæst foretages på samme vis en hypotesetest for forskellen mellem karaktergivning for kontrolgruppen og eksperimentgruppen AI bias. Da der forventes forskellige værdier for de to grupper i hver hypotesetest, tilføjes betingelsen `unequal` (Agresti, 2018: 201).

Af disse hypotesetests udledes og analyseres der først og fremmest på p-værdien samt konfidensintervallerne. Her ønskes p-værdierne at være under 0,1 for at være signifikante.

Hernæst beregnes den faktuelle samt procentuelle forskel mellem støjniveauerne for de to eksperimentgrupper i forhold til kontrolgruppen. Støjen er angivet som standardafvigelsen i stikprøven (gennemsnitlig afvigelse fra gennemsnittet) (Kahneman et al., 2021: 57; 62). Med disse beregninger klargøres det, hvorvidt projektets intervention har reduceret støjniveauet for bedømmelsen af danskopgaven.

Efterfølgende ønskes det beregnet, hvorvidt lærernes bias er reduceret. Ifølge Kahneman et al. (2021) bestemmes bias ved forskellen mellem den gennemsnitlige bedømmelse (gennemsnitskarakteren) og den sande værdi (Kahneman et al., 2021: 58). Dermed kan den faktuelle bias beregnes ved at subtrahere den gennemsnitlige karakter for kontrolgruppen, *Second Opinion Influence* og AI bias med den sande værdi (karakteren 7). Dette viser, hvor langt grupperne i gennemsnit bedømmer fejlagtigt i forhold til målet. Herefter beregnes den procentuelle forskel i bias for SOI-gruppen og AI bias-gruppen i forhold til kontrolgruppen. Med disse beregninger klargøres det, hvorvidt projektets intervention har reduceret niveauet af bias for bedømmelsen af danskopgaven.

Når standardafvigelsen er aflæst, og forskellen mellem den gennemsnitlige karakter for den respektive gruppe og den sande karakter er beregnet, kan det samlede fejlniveau beregnes. Dette sker ved følgende formel:

$$Samlede\ fejlniveau\ = Bias^2\ + Støj^2$$ (Kahneman et al., 2021: 62).

Denne formel kan med fordel omskrives til følgende formel, så den passer i projektets kontekst:

$$Samlede\ fejlniveau = (Gennemsnitskarakter - Sand\ karakter)^2\ + Standardafvigelse^2$$

MSE-værdierne for SOI-gruppen og AI bias-gruppen sammenlignes med MSE-værdien for kontrolgruppen. Desuden beregnes den procentuelle forskel i disse værdier i forhold til kontrolgruppen for at bestemme, hvor meget projektets intervention har reduceret det samlede fejlniveau for bedømmelserne af danskopgaven.

4.6.5 Undersøgelse af censorers samlede fejlniveau i karaktergivning

For at besvare projektets underhypotese (H1b) vil der foretages hypotesetests af fire omgange. Dette gøres endnu en gang ved kommandoen `ttest` i STATA. Der foretages først en hypotesetest for forskellen mellem karaktergivning for dem, der har censorerfaring i kontrolgruppen og eksperimentgruppen *Second Opnion Influence*. Dernæst foretages på samme vis en hypotesetest for forskellen mellem karaktergivning for dem, der har censorerfaring i kontrolgruppen og eksperimentgruppen AI bias. Disse to hypotesetests gentages for dem, der ikke har censorerfaring. Da der forventes forskellige værdier for de to grupper i hver hypotesetest, tilføjes betingelsen `unequal` (Agresti, 2018: 201).
Som ved første hypotesetest udledes og analyseres der først og fremmest på p-værdien samt konfidensintervallerne. Her ønskes p-værdierne at være under 0,1 for at være signifikante.

Desuden gentages tidligere beregning af den faktuelle samt procentuelle forskel mellem støjniveauerne i de to eksperimentgrupper i forhold til kontrolgruppen, men hvor forskellen i denne del af analysen er, at grupperne måles op imod hinanden i forhold til lærernes censorerfaring. Med disse beregninger klargøres det, hvorvidt *boosting* af lærerne har nogen indvirkning på, hvordan projektets intervention virker på det samlede fejlniveau for bedømmelsen af danskopgaven.

På samme vis som i forrige undersøgelse ønskes det beregnet, hvorvidt lærernes bias er reduceret, men her fordelt i forhold til censorerfaring i grupperne. Ved at udregne lærernes faktiske og procentuelle forskel i bias i mellem lærerne i forhold til censorerfaring, kan det ses, hvilken indvirkning *boosting* har på projektets intervention og brugen af beslutningsstøtteværktøjet.

Slutteligt vil hypotesen besvares efter en beregning af de samlede fejlniveau fordelt i forhold til censorerfaring i grupperne. Denne beregning sker på samme måde som ved forrige hypotese:

$$Samlede\ fejlniveau = (Gennemsnitskarakter - Sand\ karakter)^2 + Standardafvigelse^2$$

MSE-værdierne for grupperne fordelt på censorerfaring sammenlignes inden for eksperimentgrupperne og sammenlignes med resultatet i kontrolgruppen. En tolkning af disse

resultater vil på baggrund af teori kunne vise, hvordan *boosting* og mangel på samme påvirker lærernes måde at tilgå og bruge beslutningsstøtteværktøjet og dermed også, hvorledes beslutningsstøtteværktøjet virker forskelligt afhængigt af censorerfaring.

4.6.6 Undersøgelse af eksperimentgruppen AI bias' tilslutning til beslutningsstøtteværktøjet

For at besvare, hvorvidt eksperimentgruppen AI bias er mere tilbøjelig til at være uenig med den angivne feedback, genereres en klassisk frekvenstabel ved kommandoen `fre`. Denne tabel giver et hurtigt overblik over tilslutningen af feedbacken for de to eksperimentgrupper. Fordelingen af respondenterne i den ene eller den anden ende af skalaen vil sammenlignes i de to eksperimentgrupper for på den måde at klarlægge forskellen i tilslutningen mellem grupperne.

For at nuancere dette resultat vil der desuden foretages to hypotesetests for forskellen i karaktergivningen internt i de to eksperimentgrupper. Her udgør graderne for tilslutningen "Høj/Meget høj" og "Lav/Meget lav" de to grupper, som testes for forskelle imellem. Disse tests vil igen foretages indenfor et 90 procent konfidensinterval, hvorfor resultaterne godtages som signifikante, såfremt p-værdien er under værdien 0,1.

Disse hypotesetests vil afføde den nødvendige viden til at kunne bestemme det samlede fejlniveau. Dette vil foregå på samme måde som ved besvarelsen af H1a. Den eneste forskel er, at ved besvarelsen af H2 opdeles grupperne i forhold til deres tilslutning. Desuden vil der forud for bestemmelsen af det samlede fejlniveau blive tolket på både niveauet af bias og støjniveauet.

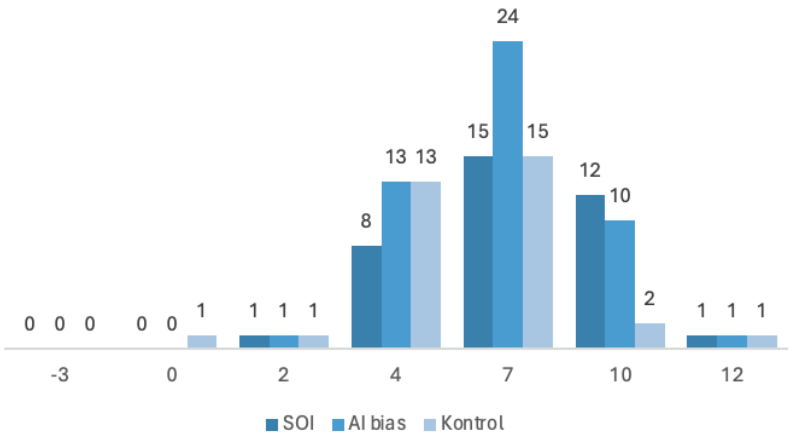
5. Analyse

Denne analyse vil være opdelt i fem dele. Der vil først redegøres for, hvor korrekte surveyeksperimentets respondenters karaktergivning er, samt hvordan dette fordeles i forhold til den sande karakter. Dernæst vil der foretages en kvalitativ analyse af lærernes egen feedback i henhold til 3.g-elevens opgavebesvarelse, som forklares med afsæt i teorien. Slutteligt følger tre analyser med henblik på besvarelse af projektets tre opstillede hypoteser. Alle dele har ét samlet formål om kummulativt at kunne besvare projektets problemformulering.

5.1 Korrektheden af karaktergivning i surveyundersøgelsen

Projektets problemformulering udspringer blandt andet af den spredning i karaktergivning i gymnasiet, som CEPOS (2021) peger på i en analyse. Her klarlægges det blandt andet, hvordan gymnasieelever generelt modtager dårligere karakterer i forbindelse med eksamen.
Det findes relevant at undersøge, hvorvidt respondenterne i projektets surveyeksperiment begår fejl i karaktergivningen.

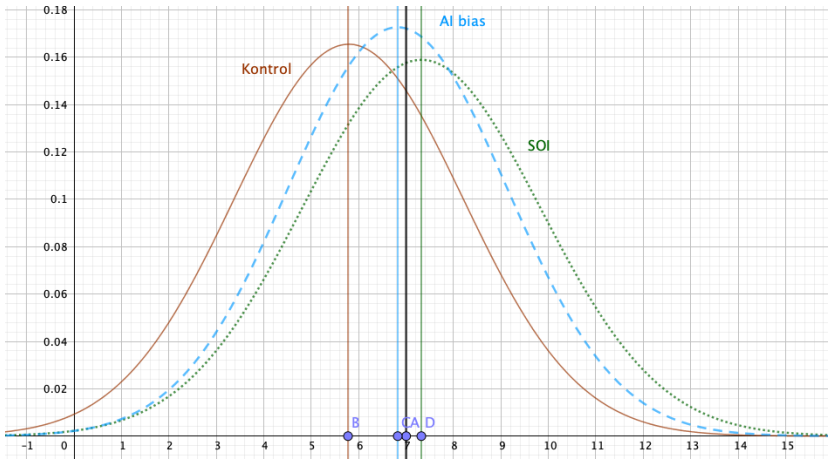
Nedenfor ses fordelingen af karakterer mellem de tre grupper i figur 12.



Figur 12: Respondenterne i surveyeksperimentets karakterfordeling fordelt i tre grupper

Her findes det først og fremmest relevant at inddrage, at de tre grupper har angivet en samlet karakter for opgavebesvarelsen med tre forskellige udgangspunkter. De to eksperimentgrupper er udsat for hver sin intervention, mens kontrolgruppen har læst og bedømt opgaven uden nogen former for intervention.

Af søjlediagrammet i figur 12 ses det tydeligt, at Kahneman et al.s (2021) teori om støj kan bakkes op af data fra surveyen. Respondenter fra alle tre grupper har angivet karakterer på alle trin fra 02 til og med 12. Dette indikerer meget tydeligt, at der er en stor spredning i karaktergivningen blandt surveyens gymnasielærere. Desuden har en enkelt respondent fra kontrolgruppen angivet karakteren 00, som betyder, at denne gymnasielærer har dumpet opgavebesvarelsen fra 3.g-eleven.
Søjlediagrammerne kan dog have den ulempe at visualisere antallet misvisende gennem absolutte tal, hvis der ikke er det samme antal respondenter i alle grupper. Derfor er Gaussfordelingerne for de tre grupper forsøgt visualiseret i Geogebra i figur 13.



Figur 13: Gaussfordelingen for respondenterne i surveyeksperimentets karakterfordeling

Figur 13 ovenfor viser Gaussfordelingerne for kontrolgruppen, SOI-gruppen og AI Bias-gruppen. X-aksen er karaktererne og y-aksen er antal respondenter. Den sorte, vertikale linje er den sande karakter (karakteren 7), som er bestemt af beslutningsstøtteværktøjet. De andre vertikale linjer er gruppernes respektive gennemsnitlige bedømmelse. Dette visualiserer endvidere, hvor langt disse karaktergennemsnit er fra den sande karakter. Interventionen har tydeligvis haft en effekt på gymnasielærerne.

Det ses til gengæld også, at spredningerne i karaktergivningen er nogenlunde ens, hvilket vidner om, at der stadig forekommer fejl i karaktergivningen i projektets surveyeksperiment trods interventionen. For yderligere at klarlægge fejlene i karaktergivningen vil der i følgende afsnit blive inddraget tekstpassager fra gymnasielærernes feedback til opgavebesvarelsen fra 3.g-eleven i projektets surveyeksperiment.

5.2 Kvalitative forklaringer

Dette afsnit vil gennemgå surveyens respondenters egne udtalelser angående opgaven såvel som feedbacken, der er genereret af AI. Mange af de lærere, som har gennemført surveyen, har sat ord på, hvordan de forholder sig til opgaven. På baggrund af disse udtalelser vil projektet gennemgå, hvorfor de i henhold til den udvalgte teori begår fejl.

Noget af det, som gymnasielærerne er meget uenige om, er sproget. En lærer skriver således: *"Den er faglig, velformuleret, reflekterende og uden nævneværdige stavefejl. Det vurderer jeg umiddelbart til et 12 tal."* (Bilag 5, Lærer 74). Denne lærer mener desuden, at fagligheden er i top. En anden lærer bakker op og finder ligeledes stort set ingen stavefejl: *"Den skriftlige fremstilling er kendetegnet ved en nogenlunde sikker syntaks og stort set ingen stavefejl."* (Bilag 5, Lærer 65).

Dette står i kontrast til mange andre lærere, der kritiserer netop sprogets lave grammatiske niveau. *"Læs korrektur, undgå sprogfejl."* (Bilag 5, Lærer 4), *"Der mangler korrekturlæsning/der er en del sproglig upræcision."* (Bilag 5, Lærer 67), *"Der er en del grammatiske fejl i dele af opgaven."* (Bilag 5, Lærer 10), *"Der er en del grammatiske sjuskefejl."* (Bilag 5, Lærer 37). Dette beviser, at lærerne ikke har en konsensus angående vægtningen af mangler og kvaliteter. De er ikke synkroniserede angående, hvad der er et lavt eller højt sprogligt niveau. Antallet af stavefejl er de samme i opgaven uanset hvilken lærer, der læser opgaven. Alligevel er der stor uenighed om, hvorvidt der er for mange eller stort set ingen grammatiske fejl. Dette vidner om, at der er stabil mønsterstøj til stede i bedømmelsespraksissen.

Et andet kritikpunkt, som lærerne er uenige om, er refleksion og den personlige stemme. Flere lærere efterspørger dette og mener, at besvarelsen bliver for analyserende: *"... Du er ikke så stærk på det dialogiske (husk at du er et jeg, med en personlig stemme, der skriver til en, der skal opfordres til at reflektere)."* (Bilag 5, Lærer 43). En anden lærer skriver omvendt følgende om 3.g-elevens opgave: *"... Fornuftig reflekterende tilgang og personlig stemme især i startfasen."* (Bilag 5, Lærer 51). Det fremgår fortsat tydeligt, at lærerne ikke selv er enige med hinanden om, hvordan en "rigtig" opgavebesvarelse skal skrives. De må derfor vurderes som afvigende for fagets reelle læringsmål og opgavens egentlige rammer. Deres personlige præferencer for, hvordan en god danskopgave skal skrives, fylder mest og dette resulterer i

yderligere mønsterstøj. Denne skævhed i deres bedømmelsespraksis resulterer yderligere i den spredning i karakter, der ses blandt surveyens respondenter.

Lærerne er ydermere meget uenige om, hvorvidt indledningen lever op til kravene inden for den reflekterende artikel. Mange af lærerne kommenterer på indledningen, og dermed tegner der sig et tydeligt mønster af, at lærerne mangler konsensus om netop denne del af opgaven. De lærere, som mener, at indledningen er fornuftig, skriver eksempelvis: *"Udmærket indledning der præsenterer emnet fokuseret og fanger læseren."* (Bilag 5, Lærer 33), og *"Du bygger godt op og har god indledning og afslutning, og også strukturen i og mellem afsnit er god og klar."* (Bilag 5, Lærer 72).

Andre lærere mener ikke, at indledningen er fokuseret nok. Eksempler på dette er følgende: *"Besvarelsen har en ufokuseret indledning."* (Bilag 5, Lærer 73), *"God, men lidt rodet indledning."* (Bilag 5, Lærer 18) og *"Derudover er indledningen meget flyvsk og kunne godt gøres lidt mere konkret."* (Bilag 5, Lærer 26). Dette vidner endnu engang om, at lærerne mangler klare retningslinjer for, hvad der er en god indledning, og hvad der er en mangelfuld indledning. Ofte er det op til lærerens egen subjektive vurdering at afgøre, om opgaven er tilfredsstillende. I dette tilfælde, hvor indledningen deler vandene, beviser det, at lærernes personlige præferencer skaber stabil mønsterstøj. Dette bidrager til det samlede fejlniveau i karaktergivningen af danskopgaven.

Desuden ses i feedbacken, at lærerne varierer i hårdheden af bedømmelserne. *"Teksterne inddrages fornuftigt.* [..] *Jeg synes, at eleven har fornuftig styr på genren. Jeg giver et flinkt 10-tal."* (Bilag 5, Lærer 51). Lærer nummer 51 giver et flinkt 10-tal, hvilket direkte vidner om en generøs bedømmelse. En anden lærer er meget hårdere i bedømmelsen: *"Fornuftig forståelse af synspunkter i hovedteksten og teksteksempler inddrages.* [...] *Læs grundigere korrektur. Hvis dette er hele afleveringen, er den ikke bestået."* (Bilag 5, Lærer 13). Disse to lærere er meget langt fra hinanden i bedømmelsen. I resterende feedback nævner de nogenlunde de samme opmærksomhedspunkter. Alligevel giver den første lærer et flinkt 10-tal, mens den anden lærer takserer manglerne til en karakter under 02. Dette er et udtryk for niveaustøj. Nogle lærere er gavmilde, mens andre straffer hårdt. Dette fører dermed yderligere til spredning i karakter for danskopgaven.

Det er nu klarlagt, at der forekommer fejl i karaktergivningen fra projektets eget surveyeksperiment. Det fremgår tydeligt af de kvalitative tekstuddrag fra lærernes feedback, at

der hersker både stabil mønsterstøj og niveaustøj i karaktergivningsprocessen. Hertil findes det relevant at undersøge, hvorvidt projektets AI-genererede beslutningsstøtteværktøj har bidraget til at mindske forekomsten af fejl i karaktergivningen.

5.3 Beslutningsstøtteværktøjets effekt på karaktergivning

Denne del af projektets analyse har til formål at besvare den primære hypotese H1a.

H1a: Gymnasielærere, som har benyttet AI som beslutningsstøtteværktøj, vil i gennemsnit lave færre fejl i karaktergivningen i forhold til kontrolgruppen.

Til at besvare denne hypotese vil der blive gjort brug af fejlligningen, der baserer sig på Gauss' metode til at bestemme det samlede fejlniveau, som ifølge Kahneman et al. (2021) bedst baseres ved *Mean Squared Error* (MSE). Udregningen af MSE er nævnt i analysestrategien afsnit 4.6.4.

Ved at sammenligne det samlede fejlniveau for eksperimentgrupperne med det samlede fejlniveau for kontrolgruppen vil det vise sig, *hvordan* og *hvor meget* interventionerne har påvirket respondenterne (Kahneman et al., 2021: 59). Herefter vil en mere dybdegående analyse pågå, der har til formål at sammenholde teorien med eksperimenternes resultater for på den måde at afklare, hvorvidt det er bias, støj eller begge dele, der er reduceret i bedømmelsesprocessen.

Som tidligere nævnt er det nødvendigt at fastslå niveauet af bias og støj i grupperne for at udregne det samlede fejlniveau (MSE) for lærerne i de tre grupper. Dermed skal gennemsnitskarakteren og standardafvigelsen i stikprøverne bestemmes.

Gennemsnitskarakteren og standardafvigelsen findes ved en hypotesetest. Denne test udregner samtidig signifikansniveau og konfidensintervallet. Følgende hypotesetests er for henholdsvis kontrolgruppen, SOI-gruppen og AI bias-gruppen, der ses i tabel 16.

Hypotesetest: Fejl i karaktergivning SOI og AI bias mod kontrolgruppe			
	Kontrol	Second Opnion Influence	AI bias
Gennemsnits-karakter	5,79	7,32	6,82
Standard-afvigelse	2,41	2,51	2,31
90%-konfidens-interval	5,08 - 6,50	6,62 - 8,02	6,26 - 7,37
P-værdi	-	0,011	0,058

Tabel 16: Hypotesetest for fejl i karaktergivning - eksperimentgrupper mod kontrolgruppe

Inden ovenstående statistiske tests aflæses med henblik på beregning af MSE, vil der tolkes på deres signifikansniveau for at vurdere repræsentativiteten. Hypotesetesten viser, at værdierne er signifikante inden for et signifikansniveau på 0,1, hvilket betyder, at der med 90 procents sandsynlighed vil kunne findes samme sammenhæng et andet tilfældigt sted i populationen. Desuden kan det ses ved samtlige konfidensintervaller, at gennemsnittet er validt, da disse konfidensintervallerne er forholdsvist snævre. Hvis konfidensintervallet er meget bredt, er validiteten af den gennemsnitlige besvarelse mindre sikker (Agresti, 2018, s. 115-117). Dette er ikke tilfældet i foregående hypotesetests.

Herefter kan værdierne aflæses, som er nødvendige for beregningen af MSE. Ses der på standardafvigelserne, som er et udtryk for spredningen af karaktererne i grupperne (og dermed støj), så er der ikke nogen nævneværdig forskel mellem grupperne.
Kontrolgruppens standardafvigelse er 2,41, hvilket vidner om en stor spredning i de givne karakterer for lærerne i denne gruppe. Som nævnt i afsnit 5.1 har lærerne i kontrolgruppen givet karakterer fra 00 til 12 for samme opgave. Dette vidner om meget støj i karaktergivningsprocessen.

SOI-gruppen har en standardafvigelse på 2,51, hvilket er højere end kontrolgruppen. Spredningen af karaktererne er dermed blevet større. Støjen er dermed vokset med 4,09 procent i forhold til kontrolgruppen, hvilket er en lille stigning.

AI bias-gruppen har en standardafvigelse på 2,31, hvilket er den laveste spredning i karakter. I forhold til kontrolgruppen er støjen reduceret med 4,21 procent, hvilket dog heller ikke er en nævneværdig forskel.

Samlet set har alle grupperne cirka den samme høje mængde støj i bedømmelsen af danskopgaven, hvilket også kan ses tydeligt i Gasussfordelingen i figur 13, da det gælder for alle grupper, at der er givet karakterer fra minimum 02 til 12. Gruppernes støjniveau og den procentuelle forskel mellem eksperimentgrupperne og kontrolgruppen ses i nedenstående tabel 17:

Ændring af støj efter intervention			
	Kontrolgruppe	Second Opnion Influence	AI bias
Støj	2,41	2,51	2,31
Ændret støj		4,09%	-4,21%

Tabel 17: Ændring af støj i karaktergivning efter intervention

At standardafvigelsen ikke har ændret sig nævneværdigt, er naturligt, da interventionen ikke har haft som primær hensigt at reducere støj. Som nævnt i teoriafsnittet er støj usynligt. Interventioner mod støj vil derfor ofte fjerne fejl forårsaget af støj, men hvilke fejl vides ikke. I denne undersøgelse har interventionen derfor ikke ændret i præcisionen af karaktergivningen. Dermed kan det siges, at lærerne i samme grad varierer i niveaustøj, stabil mønsterstøj og selvfølgelig situationsstøj som i kontrolgruppen. Dette leder naturligt op til spørgsmålet: Hvor meget bias er reduceret?

Eksperimentets intervention har været fokuseret omkring præventiv *debiasing* ved brug af *nudging* gennem *biasing information*. *Nudging* (modificering af miljøet før bedømmelse) tiltrækker effekter eller bias i en given retning for at gøre bedømmelsen mere korrekt i forhold til den sande karakter, der er angivet af AI. At lærerne får at vide hvilken karakter, beslutningsstøtteværktøjet har givet, kan have påvirket lærernes system-1-tænkning så meget, at de adopterer samme holdning.

Ud fra hypotesetestene ses det tydeligt, hvordan forskellen mellem eksperimentgruppernes gennemsnitlige karakter og den sande værdi (karakteren 7) er blevet mindre sammenlignet med kontrolgruppen.

Kontrolgruppens karaktergennemsnit er 5,79, hvilket er 1,21 karakterpoint fra 7-tallet. Denne gruppe bedømmer dermed i gennemsnit under den sande værdi, hvilket kan indikere, at deres bias er af pessimistisk karakter.

Gruppen, der er udsat for *Second Opinion Influence*, har et karaktergennemsnit på 7,32, hvilket er 0,32 karakterpoint fra karakteren 7. Dermed er denne gruppe kommet 0,89 karakterpoint tættere på målet. Dette svarer til en reducering i bias på 73,3 procent i forhold til kontrolgruppen. Til gengæld bedømmer disse respondenter en smule over målet.

Gruppen, der ved, at det er AI, der har genereret feedback og karakter, har et karaktergennemsnit på 6,82, hvilket er 0,18 karakterpoint fra karakteren 7. Dermed er denne gruppe kommet 1,03 karakterpoint tættere på målet. Dette svarer til en reducering i bias på 85 procent i forhold til kontrolgruppen. Denne gruppes bedømmelse er dermed tættest på målet af alle grupperne. Gruppernes niveau af bias og den procentuelle forskel mellem eksperimentgrupperne og kontrolgruppen ses i nedenstående tabel 18:

Ændring af bias efter intervention			
	Kontrolgruppe	Second Opnion Influence	AI bias
Gennemsnits-karakter	5,79	7,32	6,82
Bias	1,21	0,32	0,18
Ændret bias		-73,3%	-85%

Tabel 18: Ændring af bias i karakterfordeling efter intervention

Dette betyder, at interventionen har påvirket lærernes bias i den rigtige retning, forstået på den måde, at karaktergivningen er blevet mere korrekt.

Da både den gennemsnitlige karakter, den sande karakter og standardafvigelsen er fundet, kan fejlligningen benyttes til at bestemme det samlede fejlniveau for dansklærernes bedømmelse af danskopgaven (Kahneman et al., 2021: 62).

Kontrolgruppe:

$$Samlet\ fejlniveau = (5{,}787879 - 7)^2 + 2{,}407768^2 = 7{,}267\ MSE$$

SOI:

$$Samlet\ fejlniveau = (7{,}324324 - 7)^2 + 2{,}506149^2 = 6{,}386\ MSE$$

AI Bias:

$$Samlet\ fejlniveau = (6{,}816327 - 7)^2 + 2{,}306454^2 = 5{,}353\ MSE$$

Efter udregningen af gruppernes MSE ses det, hvordan det er lykkedes at reducere det samlede fejlniveau for både SOI-gruppen og AI bias-gruppen. Dette betyder, at lærerne, som har benyttet beslutningsstøtteværktøjet, i gennemsnit laver færre fejl, når de giver karakter i forhold til dem, der ikke har brugt beslutningsstøtteværktøjet.

Det samlede fejlniveau er faldet 0,881 MSE (12,1 procent) for SOI-gruppen og 1,914 MSE (26,3 procent) for AI bias-gruppen. Dette betyder, at gymnasielærere kan mindske antallet og ikke mindst størrelsen på fejl i karaktergivning i skriftlig dansk ved at bruge AI som beslutningsstøtteværktøj. Dette ses i nedenstående tabel 19:

Opsamling på fejlniveauer			
	Kontrolgruppe	Second Opnion Influence	AI bias
Samlet fejlniveau	7,27 MSE	6,39 MSE	5,35 MSE
Ændret fejlniveau		-12,1%	-26,3%

Tabel 19: Opsamling på fejlniveauer for karakterfordeling

Dermed kan projektets primære hypotese bekræftes. Når gymnasielærerne benytter beslutningsstøtteværktøjet, reduceres det samlede fejlniveau i karaktergivningsprocessen og dermed har interventionen ved *debiasing*, herunder *nudging* og påvirkning gennem *biasing information,* virket efter hensigten.

5.4 Har censorer et lavere samlet fejlniveau?

Kahneman et al. (2021) nævner flere forskellige måder, hvorpå *debiasing* kan foregå. En anden måde end *nudging* er *boosting*. Som nævnt i afsnit 3.1.6 kan fagprofessionelle trænes til at blive opmærksomme på deres egen bias. Dette virker dog kun inden for det specifikke fagområde. Generelt har intelligens, uddannelse og akademisk niveau også, ifølge teorien, en positiv sammenhæng mellem karaktergivning og korrekthed. Som nævnt i problemfeltet, tilbydes censorer vejledning og kurser i forbindelse med bedømmelse af opgaver. Dermed må det antages, at disse gymnasielærere har et højere fagligt niveau, når det kommer til at bedømme opgaver i forhold til lærere, der ikke er censorer, og derfor ikke har fået vejledning.
Dette leder naturligt op til spørgsmålet, som H1b baseres på; er der forskel i fejlniveauet for de lærere, der har været eller er censor i forhold til dem, der aldrig har været censor?

H1b: Gymnasielærere, der er eller har været censorer og som anvender beslutningsstøtteværktøjet, vil have et lavere samlet fejlniveau i forhold til de lærere, der ikke har censorerfaring

For at kunne beregne det samlede fejlniveau for ikke-censorer sammenlignet med censorer, så vil gennemsnitskarakteren og standardafvigelsen igen findes ved en hypotesetests. Følgende hypotesetests er for henholdsvis kontrolgruppen, SOI-gruppen og AI bias-gruppen, men hvor grupperne er opdelt i censorstatus. Disse hypotesetests resultater ses i nedenstående tabel 20:

Hypotesetest: Er/har været censor SOI og AI bias mod kontrolgruppe				Hypotesetest: Har aldrig været censor SOI og AI bias mod kontrolgruppe			
	Kontrol	Second Opnion Influence	AI bias		Kontrol	Second Opnion Influence	AI bias
Gennemsnits-karakter	5,78	7,04	6,82	Gennemsnits-karakter	5,83	7,85	6,8
Standard-afvigelse	2,26	2,56	2,32	Standard-afvigelse	3,25	2,41	2,35
90%-konfidens-interval	5,03 - 6,52	6,15 - 7,94	6,10 - 7,56	90%-konfidens-interval	3,16 - 8,51	6,65 - 9,04	5,89 - 7,71
P-værdi	-	0,070	0,091	P-værdi	-	0,214	0,521

Tabel 20: Hypotesetests for censor versus ikke-censorer i de tre grupper

Inden ovenstående statistiske tests aflæses med henblik på beregning af MSE, vil der tolkes på deres signifikansniveau for at vurdere repræsentativiteten. Hypotesetestene er signifikante på

et signifikansniveau på 0,1 ved de to hypotesetests, der sammenligner censorer i alle tre grupper, men ikke for dem, der ikke har været censorer.
For hypotesetestene kun for censorer gælder det, at der med 90 procent sandsynlighed vil kunne findes samme sammenhæng et andet tilfældigt sted i populationen. Dette er ikke tilfældet ved hypotesetestene for gymnasielærere, der ikke har censorerfaring, da p-værdierne for SOI-gruppen og AI bias-gruppen er henholdsvis 0,214 og 0,521. Disse resultater vil dog fortsat indgå i analysen, men med et forbehold for repræsentativiteten. De lave signifikansniveauer kan skyldes det lave antal af gymnasielærere uden censorerfaring, som indgår i stikprøven. For disse hypotesetests gælder det desuden, at konfidensintervallerne er brede, hvilket kan indikere, at validiteten af gennemsnittet er mindre sikker (Agresti, 2018, s. 115-117). Dette er dog ikke gældende for hypotesetestene kun med censorer.

Herefter kan værdierne aflæses, som er nødvendige for beregningen af MSE. Ser man på standardafvigelserne, som er et udtryk for spredningen af karaktererne i grupperne (og dermed støj), så er der ikke et tydeligt mønster mellem grupperne og heller ikke mellem censor og ikke-censor.
Gymnasielærerne i kontrolgruppen, der bedømmer opgaver, som de plejer uden brug af beslutningsstøtteværktøjet, støjer 30,5 procent mindre, når de har censorerfaring i forhold til de gymnasielærere i kontrolgruppen, der ikke har censorerfaring. Dette passer fint på hypotesen og Kahneman et al.s (2021) tese om, at *boosting* reducerer fejl (Kahneman et al., 2021: 238).

Forskellen bliver til gengæld mindre, når lærerne benytter beslutningsstøtteværktøjet. For SOI-gruppen gælder det, at gymnasielærerne med censorerfaring i gennemsnit støjer 4,2 procent mere end dem, der ikke har censorerfaring. For AI bias-gruppen gælder det, at gymnasielærerne med censorerfaring støjer 1,3 procent mindre end dem, der ikke har censorerfaring. Disse resultater ses i nedenstående tabel 21:

Støj og censorstatus						
	Kontrolgruppe		Second Opnion Influence		AI bias	
	Ikke censor	Censor	Ikke censor	Censor	Ikke censor	Censor
Støj	3,25	2,26	2,41	2,56	2,35	2,32
Forskel i støj	-	-30,5%	-	4,2%	-	-1,3%

Tabel 21: Støj i karaktergivning og censorstatus

Tabel 21 viser, at den *boosting*, som censorerne har fået, bliver mindre væsentlig, da de gymnasielærere, der ikke har censorerfaring har et støjniveau, der minder meget om støjniveauet for censorer. I henhold til Kahneman et al. (2021) kan dette tyde på, at lærerne uden censorerfaring bliver hjulpet så meget af beslutningsstøtteværktøjet, at deres stabile mønsterstøj og niveaustøj bliver mindsket i karaktergivningsprocessen.

Som i den forrige udregning af det samlede fejlniveau er det nærtliggende at beregne niveauet af bias for eksperimentgrupperne, i forhold til om gymnasielærerne har censorerfaring eller ej. Både for gymnasielærerne i SOI-gruppen og i AI bias-gruppen gælder det, at gymnasielærere uden censorerfaring i gennemsnit bedømmer længere fra den sande karakter i forhold til dem, der har censorerfaring. I SOI-gruppen er bias ved dem, der har censorerfaring, 95,3 procent mindre end dem, der ikke har censorerfaring. Niveauet af bias i AI bias-gruppen er 15 procent mindre hos gymnasielærerne med censorerfaring versus dem uden. Dog skal det pointeres, at både censorer og ikke-censorer i begge eksperimentgrupper er tættere på den rigtige karakter i forhold til kontrolgruppen.

Forskellen mellem censorerne i SOI-gruppen og AI bias-gruppen er også bemærkelsesværdig. Når censorerne ved, at det er AI, der har bedømt opgaven, falder den gennemsnitlige karakter, og mængden af bias bliver større. Faktisk har censorerne i AI bias-gruppen en bias, der er 325 procent større end censorerne i SOI-gruppen. Dette kan ifølge Kahneman et al.s (2021) teori tolkes som objektiv ignorance, der har en negativ påvirkning på de fagprofessionelles holdning til trusler mod det fagprofessionelle skøn, som i dette tilfælde er AI. Denne objektive ignorance

etablerer fordomme om AI, der aktiverer konklusionsbias og dermed modsætter beslutningsstøtteværktøjets vurdering i højere grad. Dette fremgår af nedenstående tabel 22:

Bias og censorstatus						
	Kontrolgruppe		Second Opnion Influence		AI bias	
	Ikke censor	Censor	Ikke censor	Censor	Ikke censor	Censor
Karakter-gennemsnit	5,83	5,78	7,85	7,04	6,8	6,83
Bias	1,17	1,22	0,85	0,04	0,2	0,17
Forskel i bias	-	4,3%	-	-95,3%	-	-15%

Tabel 22: Bias i karaktergivning og censorstatus

Da både den gennemsnitlige karakter, den sande karakter og standardafvigelsen er fundet for både kontrolgruppen og begge eksperimentgrupper fordelt i forhold til om gymnasielæreren har censorerfaring eller ej, kan fejlligningen benyttes til at bestemme det samlede fejlniveau for disse grupper (Kahneman et al., 2021: 62).

Kontrolgruppe:

Censor: $Samlet\ fejlniveau = (5{,}777778 - 7)^2 + 2{,}258886^2 = 6{,}596$ MSE

Ikke censor: $Samlet\ fejlniveau = (5{,}833333 - 7)^2 + 3{,}250641^2 = 11{,}928$ MSE

SOI:

Censor: $Samlet\ fejlniveau = (7{,}041667 - 7)^2 + 2{,}561914^2 = 6{,}565$ MSE

Ikke censor: $Samlet\ fejlniveau = (7{,}846154 - 7)^2 + 2{,}409915^2 = 6{,}524$ MSE

AI Bias:

Censor: $Samlet\ fejlniveau = (6{,}827586 - 7)^2 + 2{,}315614^2 = 5{,}392$ MSE

Ikke censor: $Samlet\ fejlniveau = (6{,}8 - 7)^2 + 2{,}35305^2 = 5{,}577$ MSE

Efter udregningen af gruppernes MSE fordelt på censorstatus, må det siges, at resultaterne er tvetydige. Der er stor forskel på fejlniveauet i kontrolgruppen i forhold til om læreren har censorerfaring eller ej. Dog skal det nævnes, at data fra hypotesetestene angående lærere uden censorerfaring ikke er signifikante. Derfor kan denne forskel være misvisende. Alligevel er det slående, hvor stor forskel i det samlede fejlniveau, der er mellem censorerne og ikke-censorerne. I kontrolgruppen har de lærere, der har censorerfaring, et samlet fejlniveau, der er 44,7 procent mindre end dem, der ikke har censorerfaring.

Når der foretages samme beregninger for begge eksperimentgrupper, er forskellene langt mindre. Som forrige beregninger viser, har AI bias-gruppen det laveste samlede fejlniveau, hvor der ikke er nogen nævneværdig forskel mellem censor og ikke-censor. Sidstnævnte gælder også SOI-gruppen.

Dette er til gengæld bemærkelsesværdigt, da der kunne forventes en lignende sammenhæng mellem censorer og ikke-censorer, som i kontrolgruppen. Dette er ikke tilfældet, hvilket kan tolkes på forskellige måder.

Beslutningsstøtteværktøjet kan hjælpe de lærere, der ikke er censorer, så meget, at de kan reducere deres fejlniveau, så det kommer på samme lave niveau, som censorernes fejlniveau. Dermed bliver *boosting* af gymnasielærere, der skal udføre en censorgerning, mindre vigtigt, da gymnasielæreren nu i stedet kan læne sig op af beslutningsstøtteværktøjet. Lærere, som ikke har censorerfaring kan dermed ved brug af beslutningsstøtteværktøjet, give karakterer, som er lige så korrekte som dem, der har censorerfaring samt har fået kurser og dermed vejledning i bedømmelser af danskopgaver.

Dertil kan det tolkes på resultaterne, at censorerne bedømmer som de plejer, men til gengæld undervurderer opgavens niveau en lille smule for at modsætte sig beslutningsstøtteværktøjets vurdering som følge af den objektive ignorance.

Samlet fejlniveau og censorstatus						
	Kontrolgruppe		Second Opnion Influence		AI bias	
	Ikke censor	Censor	Ikke censor	Censor	Ikke censor	Censor
Samlet fejlniveau	11,93 MSE	6,60 MSE	6,52 MSE	6,57 MSE	5,58 MSE	5,39 MSE
Forskel i fejlniveau	-	-44,7%	-	0,8%	-	-3,4%

Tabel 23: Samlet fejlniveau for karaktergivning og censorstatus

Den kvantitative analyse, der har til formål at besvare hypotese H1b, har påvist, at denne hypotese kan afkræftes. Beslutningsstøtteværktøjet hjælper ikke gymnasielærere med censorerfaring til at få et endnu lavere fejlniveau. I stedet er det snarere de lærere, der ikke har censorerfaring, der bliver hjulpet så meget, at disse foretager bedømmelser og giver karakter med et lige så lavt fejlniveau som de lærere, der har modtaget *boosting*. Ydermere viser resultaterne, at censorerne i AI bias-gruppen sandsynligvis modsætter sig beslutningsstøtteværktøjet en smule på grund af den objektive ignorance jævnfør Kahneman et al. (2021), der er udfoldet i afsnit 3.1.12.

På baggrund af nysgerrigheden om, hvorvidt den fagprofessionelle stolthed i erhvervet kan påvirke tilslutningen og opfattelsen af brugbarheden af beslutningsstøtteværktøjet, er det nærtliggende at undersøge dette i en lignende undersøgelse i H2.

5.5 Eksperimentgruppernes tilslutning til beslutningsstøtteværktøjet

Anden del af dette projekts analyse har til formål at besvare den opstillede hypotese H2.

H2: Eksperimentgruppen AI bias vil være mere tilbøjelig til at være uenige med feedback- og karaktergivning genereret af AI i forhold til eksperimentgruppen Second Opinion Influence.

Fordeling af de to eksperimentgruppers tilslutning til den angivne feedback		
	Second opnion influence	AI bias
Meget høj / høj	46,5%	39%
Meget lav / lav	53,5%	61%

Tabel 24: De to eksperimenters tilslutning til den angivne feedback - andele

Ovenstående tabel 24 er en frekvenstabel over de to eksperimentgruppers tilslutning til den angivne feedback. Kategorien "Hverken/eller" er kodet missing, hvorfor de to gruppers størrelse naturligt bliver mindre.

Det er bemærkelsesværdigt, at gruppen *Second Opinion Influence* er så godt som ligeligt delt i to. 53,5 procent af respondenterne i denne gruppe kan i meget lav til lav grad tilslutte sig den givne feedback, mens 46,5 procent af respondenterne i meget høj til høj grad kan tilslutte sig den angivne feedback.

I AI bias-gruppen er respondenterne præsenteret for, at den angivne feedback er genereret af kunstig intelligens. Her kan 61 procent af respondenterne i meget lav til lav grad tilslutte sig den angivne feedback, mens 39 procent af respondenterne i denne gruppe omvendt kan tilslutte sig den AI-genererede feedback i høj til meget høj grad.

Dette resultat ræsonnerer med Kahneman et al. (2021), der netop understreger, at eksperter og fagprofessionelle på forhånd ofte vil være mistroiske overfor algoritmer og kunstig intelligens (Kahneman et al., 2021: 145). Det er dog vigtigt at pointere, at dette er gældende for netop fagprofessionelle. Som nævnt i afsnit 3.1.12 er almindelige mennesker, ifølge Kahneman et al. (2021), ikke mistænkelige over for AI. Folk vil faktisk ofte tage et råd af AI frem for af et menneske, men kun indtil det tidspunkt, hvor AI laver en fejl. Herefter falder tilliden markant (Kahneman et al., 2021: 135). AI er ikke perfekt og vil lave fejl. Dette afføder den objektive ignorance, der altid vil sætte en stopper for, at AI erstatter menneskelig dømmekraft (Kahneman et al., 2021: 146). Mennesker vil generelt aldrig vælge et alternativ og ændre deres adfærd, medmindre alternativet er nær-perfekt (Kahneman et al., 2021: 135). Det betyder, at

de lærere, som ved, at det er AI, der har genereret feedbacken, automatisk vil have en fordom om AI. Fordommen aktiverer system-1-tænkningen, som er den første hurtige intuitive tanke om noget. Allerede her hælder læreren til en side: "Kan jeg lide det, eller kan jeg ikke lide det?" Herefter vil der i mange tilfælde aktiveres konklusionsbias, hvor der enten springes direkte til konklusionen, eller hvor system-2-tænkningen mobiliseres til at matche den første intuitive system-1-tænkning (Kahneman et al., 2021: 169). Dermed har gymnasielæreren allerede en forvrænget bias.

Lærernes bias kan desuden ses ved placeringen af den gennemsnitlige karakter i forhold til den sande karakter. Både SOI-gruppen og AI bias-gruppen har reduceret det samlede fejlniveau, men AI bias-gruppen har et karaktergennemsnit, som er under den sande karakter, hvilket på baggrund af TAM, kan tolkes som en skepsis og pessemisme til beslutningsstøtteværktøjets brugbarhed. TAM beskæftiger sig netop med lærerens kognitive proces. Lærerne vurderer modellens output og sammenligner det med deres egen vurdering. Her forekommer der en negativ ladet respons, og teknologien må derfor forstås som ikke brugbar, hvoraf læreren ikke har intentioner om at anvende beslutningsstøtteværktøjet. Ydermere er der også en effekt fra lærernes sociale indflydelse. Villigheden er ligetil, da der for læreren ikke er konsekvenser for ikke at følge outputtet. Dog antyder resultatet, at den subjektive norm såvel som lærerens image er i spil i den samlede tilslutning. En lærer vil nødigt fremstå forkert over for hverken sine kollegaer eller eleverne. Derfor er den rette etos nødvendig. Hvis kollegerne ikke kan se sig enig i et individs holdning til brug af AI som beslutningsstøtteværktøj, vil læreren være mere tilbøjelig til at frasige sig det for at bibeholde ens sociale og/eller faglige status.

Dermed er det nærtliggende at undersøge, hvorvidt respondenterne i AI bias-gruppen, der ikke kan tilslutte sig den AI-genererede feedback, vil have et højere fejlniveau end de lærere, der kan tilslutte sig feedbacken. Derfor beregnes det samlede fejlniveau (MSE). For at udregne det samlede fejlniveau for lærerne i de to eksperimentgrupper er det nødvendigt at fastslå niveauet af bias og støj i grupperne fordelt på tilslutningen til feedbacken. Dermed skal gennemsnitskarakteren og standardafvigelsen i stikprøverne bestemmes.

Gennemsnitskarakteren og standardafvigelsen findes ved en hypotesetest. Denne test udregner samtidig signifikansniveau og konfidensintervallet. Følgende hypotesetests er for henholdsvis SOI-gruppen og AI bias-gruppen fordelt på tilslutningsen til feedback, der ses i tabel 25:

Hypotesetests: Fejl i karaktergivning og tilslutning til beslutningsstøtteværktøj				
	Second Opnion Influence		AI bias	
	Meget høj / høj	Meget lav / lav	Meget høj / høj	Meget lav / lav
Gennemsnits-karakter	7,92	6,6	7,64	6,18
Standard-afvigelse	1,44	3,14	1,28	2,61
90%-konfidens-interval	7,21 - 8,64	5,17 - 8,03	7,04 - 8,25	5,22 - 7,14
P-værdi	0,158		0,032	

Tabel 25: Hypotesetests for fejl i karaktergivning og tilslutning til undersøgelsens beslutningsstøtteværktøj

Før beregning af MSE vurderes hypotesetestenes signifikansniveau for at bedømme deres repræsentativitet. Hypotesetestene viser signifikans på et 0,1 niveau for eksperimentgruppen AI bias, men ikke for SOI-gruppen. For AI bias-gruppen betyder dette, at der med 90 procent sandsynlighed kan findes en lignende sammenhæng et andet tilfældigt sted i populationen. Dette gælder ikke for SOI-gruppen, da dens p-værdi er 0,1581. Resultaterne fra begge grupper vil dog indgå i analysen, men med et forbehold for repræsentativiteten. For SOI-gruppen er konfidensintervallerne også brede, hvilket kan indikere mindre sikkerhed i validiteten af gennemsnittet (Agresti, 2018: 115-117). Dette er ikke tilfældet for AI bias-gruppen.
Herefter kan værdierne aflæses, som er nødvendige for beregningen af MSE. Ses der på standardafvigelserne, så er der stor forskel i denne i forhold til om respondenten kan tilslutte sig feedbacken eller ej.

For SOI-gruppen gælder det, at standardafvigelsen er 1,44 for dem, der kan tilslutte sig feedbacken, mens standardafvigelsen er 3,14 for dem, der ikke kan tilslutte sig feedbacken. Dermed støjer de lærere, der ikke er enige i feedbacken, 118,06 procent mere end dem, der er enige. Dette er ikke overraskende, da kun 20 procent af gymnasielærerne i SOI-gruppen, som ikke kan tilslutte sig feedbacken, har givet danskopgaven karakteren 7 jævnfør tabel 26. Desuden har opgaven fået alle karakterer fra 02 til 12. Omvendt ser det ud for dem, der kan

tilslutte sig feedbacken. Her har opgaven kun fået karakteren 7 og 10, hvor karakteren 7 blev givet i 69,23 procent af tilfældene.

Karakterfordeling og tilslutning				
	Second Opnion Influence		AI bias	
	Meget høj / Høj	Meget lav / Lav	Meget høj / Høj	Meget lav / Lav
12	-	6,7%	-	4,6%
10	30,8%	26,7%	21,4%	13,6%
7	69,2%	20%	78,6%	36,4%
4	-	40%	-	40,9%
02	-	6,7%	-	4,6%

Tabel 26: Karaktergivning for eksperimentgrupper fordelt på tilslutning

For AI bias-gruppen ses samme mønster. Standardafvigelsen er 1,28 for dem, der er enige i feedbacken, mens den er 2,61 for dem, der ikke er enige. Dette udgør en procentvis forskel på 104,47 procent, hvilket er mindre end forskellen i SOI-gruppen, men stadig betydelig. Dette resultat er ikke overraskende, da kun 36,36 procent af lærerne i AI bias-gruppen, der er uenige i feedbacken, har givet danskopgaven karakteren 7. Desuden har opgaven fået alle karakterer fra 02 til 12 i denne gruppe. For dem, der er enige i feedbacken, har opgaven kun fået karaktererne 7 og 10, hvoraf der i 78,57 procent af tilfældene blev givet karakteren 7.

Dermed kan det tolkes at tilslutningen til feedbacken og dermed den opfattede brugbarhed, må have stor indvirkning på støjniveauet i bedømmelsen ved brugen af beslutningsstøtteværktøjet. Dette vil blive elaboreret senere i afsnittet. Støjniveauerne for både SOI-gruppen og AI bias-gruppen, samt den procentuelle forskel i forhold til tilslutning, kan ses i nedenstående figur 27:

Støj og tilslutning til feedback				
	Second Opnion Influence		AI bias	
Tilslutning til feedback	Meget høj / høj	Meget lav / lav	Meget høj / høj	Meget lav / lav
Støj	1,44	3,35	1,28	2,61
Forskel i støj		132,5%		104,5%

Tabel 27: Støj i karaktergivning for eksperimentgrupper fordelt på tilslutning

Som i besvarelsen af H1a er det nærtliggende at beregne niveauet af bias for eksperimentgrupperne i forhold til om de kan tilslutte sig feedbacken eller ej.

Ud fra hypotesetestene ses det, at der er forskel i bias afhængigt af om lærerne kan tilslutte sig feedbacken eller ej. Dog er der ikke et tydeligt mønster.

SOI-gruppen har et karaktergennemsnit på 7,92 for dem, der kan tilslutte sig feedbacken og et gennemsnit på 6,6 for dem, der ikke kan tilslutte sig feedbacken. Dermed er der stor forskel på den gennemsnitlige besvarelse. Dog er det overraskende nok dem, der ikke kan tilslutte sig feedbacken, der er mindst biased i forhold til dem, der kan tilslutte sig feedbacken. Disse er 56,67 procent tættere på karakteren 7.

AI bias-gruppen har et karaktergennemsnit på 7,64 for dem, der kan tilslutte sig feedbacken og et gennemsnit på 6,18 for dem, der ikke kan tilslutte sig feedbacken. Dermed er der igen stor forskel på den gennemsnitlige besvarelse. Her er det til gengæld dem, der kan tilslutte sig feedbacken, der er mindst biased i forhold til dem, der ikke kan tilslutte sig feedbacken. Disse er 27,27 procent tættere på karakteren 7. Dermed er det svært at sige noget generelt om mønsteret for tilslutning til feedback og bias. Mønsteret i denne sammenhæng kan til gengæld tolkes således at dem, der er enige, i gennemsnit bedømmer over karakteren 7, mens dem, der ikke er enige, i gennemsnit bedømmer under karakteren 7. Forskelle i niveauet af bias og den procentuelle forskel mellem eksperimentgrupperne fordelt på tilslutning, kan ses i nedenstående tabel 28:

Bias og tilslutning til feedback				
	Second Opnion Influence		AI bias	
Tilslutning til feedback	Meget høj / høj	Meget lav / lav	Meget høj / høj	Meget lav / lav
Gennemsnits-karakter	7,92	6,6	7,64	6,18
Bias	0,92	0,4	0,64	0,82
Forskel i bias		-56,7%		27,3%

Tabel 28: Bias i karaktergivning for eksperimentgrupper fordelt på tilslutning

Da både den gennemsnitlige karakter, den sande karakter og standardafvigelsen er fundet for begge eksperimentgrupper fordelt i forhold til tilslutning til feedbacken, kan fejlligningen benyttes til at bestemme det samlede fejlniveau for disse grupper (Kahneman et al., 2021: 62).

SOI:

Lav tilslutning: $Samlet\ fejlniveau = (6{,}6 - 7)^2 + 3{,}135055^2 = 9{,}989$ MSE

Høj tilslutning: $Samlet\ fejlniveau = (7{,}923077 - 7)^2 + 1{,}441153^2 = 2{,}929$ MSE

AI Bias:

Lav tilslutning: $Samlet\ fejlniveau = (6{,}181818 - 7)^2 + 2{,}611994^2 = 7{,}492$ MSE

Høj tilslutning: $Samlet\ fejlniveau = (7{,}642857 - 7)^2 + 1{,}277446^2 = 2{,}045$ MSE

Efter udregningen af de opdelte eksperimentgruppers MSE fordelt på om lærerne kan tilslutte sig feedbacken eller ej, ses det, at dem, der ikke kan tilslutte sig, har et langt højere fejlniveau end dem, der kan tilslutte sig feedbacken. Som det fremgår af tabel 29, gælder det for SOI-gruppen, at dem, der ikke kan tilslutte sig, har et samlet fejlniveau, der er 241 procent højere end dem, der kan tilslutte sig. I AI bias-gruppen gælder det, at dem, der ikke kan tilslutte sig, har et samlet fejlniveau, der er 266,4 procent højere end dem, der kan tilslutte sig. Dette betyder, at gymnasielærere, der kan tilslutte sig feedbacken mindsker antallet og ikke mindst

størrelsen på fejl i karaktergivning i skriftlig dansk ved at bruge AI som beslutningsstøtteværktøj.

Opsamling på fejlniveauer: Tilslutning til feedback i eksperimentgrupper				
	Second Opnion Influence		AI bias	
Tilslutning til feedback	Meget høj / høj	Meget lav / lav	Meget høj / høj	Meget lav / lav
Samlet fejlniveau	2,93 MSE	9,99 MSE	2,05 MSE	7,49 MSE
Forskel i fejlniveau		241%		266,4%

Tabel 29: Samlede fejlniveauer for karaktergivning i eksperimentgrupper fordelt på tilslutning

Det findes desuden interessant at se på den gruppe, der har det laveste fejlniveau af alle grupperne. Dette er den andel af AI bias-gruppen, der kan tilslutte sig feedbacken. Gruppen har et samlet fejlniveau på 2,05 MSE. Dette resultat giver god mening, da lærerne netop finder teknologiens output brugbart, hvorfor deres egen karakter naturligt vil komme tættere på den sande værdi. Denne andel af gruppen giver, som det fremgår af tabel 28, en gennemsnitskarakter på 7,64. Som det fremgår i tabel 26, giver 78,6 procent af gruppen karakteren 7 og de resterende 21,4 procent giver karakteren 10. Sammenlignes det samlede fejlniveau for dem, der kan tilslutte sig feedbacken med det samlede fejlniveau for hele AI bias-gruppen i tabel 19, viser det sig, at det er muligt, at reducere det samlede fejlniveau fra 5,35 MSE til 2,05 MSE. Dette betyder, at når gymnasielærerne i AI bias-gruppen, der kan tilslutte sig feedbacken og dermed opfatter sprogmodellen som brugbar, vil det samlede fejlniveau sænkes med 61,8 procent i forhold til hele AI bias-gruppen.

Hertil forekommer det naturligt at undersøge, hvorfor lærerne ikke opfatter karakter og feedback genereret af kunstig intelligens som brugbar. Da lærerne ikke direkte er blevet spurgt om, hvorfor de fravælger beslutningsstøtteværktøjet, vil projektet gennemgå de besvarelser, lærerne selv har givet. Her vil der tages udgangspunkt i de respondenter i AI bias-gruppen, der har en meget lav til lav grad af tilslutning. Flere af de lærere, der har opgivet lav eller meget

lav grad af tilslutning til karakter og feedback, skriver, at de ikke synes, der er sammenhæng mellem den givne opgave og modellens output. De hæfter sig også ved, at modellen i deres optik har misforstået opgavens skrivegenre: "*... og det virker også lidt som om, at AI ikke har styr på kravene til genren, når den beder om mere analyse i en reflekterende artikel"* (Bilag 5, Lærer 89). En anden lærer beretter: *"Min feedback er ikke så struktureret som AI's og heller ikke så pædagogisk fordi jeg ikke kan bruge mere tid på den. Til gengæld er den ikke en eklatant sludder for en sladder"* (Bilag 5, Lærer 93). Der er altså ifølge TAM et problem med kvaliteten af det output, AI har genereret.

Kvaliteten af beslutningsstøtteværktøjets output er samtidig også tæt forbundet med jobrelevans. Lærerne bruger i forløb med deres kognitive proces det fagprofessionelle skøn til at vurdere kvaliteten. De har vurderet, at teknologien ingen relevans har for deres daglige arbejde og finder den derfor ikke brugbar. Dette resulterer da i et samlet fejlniveau på 7,49 for gruppen, der ikke finder teknologien brugbar jævnfør tabel 29. Det er en stigning i det samlede fejlniveau på 266,4 procent i forhold til gruppen, der mener teknologien er brugbar. Det er beskrevet tidligere i projektet, at grundet sprogmodellens neurale netværk og usynlige lag, vil der være data, som det ikke er muligt at redegøre for. Der vil derfor være mulighed for, at modellens output er baseret på andet end læreplanen for dansk på A-niveau, som den er instrueret til. Det må derfor kunne udledes, at hvis det var muligt at programmere en sprogmodel, hvis usynlige lag bestod udelukkende af de danskfaglige krav og mål, ville det kunne hæve den opfattede brugbarhed blandt gymnasielærerne. Hermed vil problematikken omhandlende generativ kunstig intelligens' sorte boks være fjernet, og det er muligt at se, hvad modellens output er skabt af.

Omvendt for gruppen der har en høj til meget høj grad af tilslutning til modellens output, er der intet af deres feedback til opgaven, der er direkte uenig med det, som fremgår af feedbacken fra kunstig intelligens. Hvor kvaliteten af outputtet og jobrelevans var afgørende for den anden gruppes opfattede brugbarhed, er der for denne gruppe ingen tegn på en direkte indvirkning fra de eksterne faktorer i TAM. Det vurderes derfor, at de ved deres egen vurdering af opgaven har anvendt beslutningsstøtteværktøjet, hvilket har resulteret i, at deres samlede karakter ligger så tæt på den sande karakter. Den opfattede brugbarhed i TAM er i dette tilfælde positiv, hvilket resulterer i et markant reduceret fejlniveau. Desuden ses det, at lærernes bias flytter sig til højre på karakterskalaen, hvilket betyder, at lærerne i gennemsnit angiver karakteren over den sande værdi.

Nogle af lærerne i gruppen, der har en lav eller meget lav grad af tilslutning til feedbacken, ender faktisk med at give en lignende kritik af opgaven. Som eksempel ønsker en af lærerne mere refleksion end gennemgang af de udvalgte tekster:

"*Velformuleret opgave som lever op til genrekravene. Du kommer omkring samtlige tekster og forholder dig reflekterende til dem. Der udvises danskfaglighed løbende, uden at det bliver en analyserende artikel. En bredere perspektivering ønskes for at højne opgavens niveau."* (Bilag 5, Lærer 103)

Trods at læreren har en lav grad af tilslutning til feedbacken, der er genereret af kunstig intelligens, er denne vurdering overvældende enig med sprogmodellens. Feedbacken fra kunstig intelligens noterer opgavens styrker som: "*Du har en klar og velformuleret introduktion. Du har inddraget relevante citater fra tekstuddragene*" (Bilag 4). Lærerens egen vurdering læner sig meget op af den, som er genereret af kunstig intelligens. Det må derfor kunne fortolkes, at nogle af lærerne benægter sig at anvende beslutningsstøtteværktøjet udelukkende fordi, at den er genereret af kunstig intelligens. Det kan der være flere årsager til. Den eksterne effekt i TAM, villighed, kunne have en indvirkning på dette, da der ikke er nogen direkte konsekvens af ikke at benytte teknologien. I projektets survey bliver sprogmodellens output blot præsenteret, og der følger ingen krav om at anvende det den skriver i sin vurdering af opgaven. Den hurtige frakendelse af vurderingen kan også være grundet lærernes egen bias mod kunstig intelligens. Da denne gruppe lærere er i gruppen AI-bias, får de som det første at vide, at den vurdering de skal se, er dannet af kunstig intelligens. Derfor vil deres fordomme komme i forkøbet, og de vil derfor anvende system-1 tænkning, hvilket resulterer i, at de ikke kan tilslutte sig feedbacken af den ene årsag, at den er skrevet af kunstig intelligens, og ikke fordi det den skriver er fagligt forkert. Lærernes associationer med kunstig intelligens - og særligt sprogmodeller eller chatbots - må i dette tilfælde anses som negative. Denne bias til kunstig intelligens resulterer i et højt samlet fejlniveau jævnfør tabel 29. Dette kan også være grundet lærernes subjektive norm og image. Dog er dette projekts surveyundersøgelse anonym, og respondenterne skal derfor ikke stå til ansvar for nogen på baggrund af deres svar. Men det må stadig kunne forventes, at lærerne er blevet udfordret på deres sociale status, når de stilles over for kunstig intelligens. Som beskrevet i problemfeltets afsnit 1.3 har kunstig intelligens og chatbots fyldt meget i skolesystemet i de seneste år; særligt i form af snyd og plagiat i forbindelse med elevernes opgaver (Romme-Mølby, 2023). På baggrund af denne viden sammenholdt med resultaterne i dette afsnit, tyder det på, at snakken om kunstig intelligens på

lærerværelset har været i en negativ tone, hvilket kan have ført til negative holdninger om selvsamme problemstilling for gymnasielærerne. Det er derfor muligt, at de gymnasielærere, der har skulle tage stilling til feedback skrevet af kunstig intelligens, vil være underlagt de nærmeste kollegers holdninger, der i denne situation anses som at være imod kunstig intelligens i skolesystemet. Dette vil da resultere i, at lærerens egen fagprofessionelle vurdering ikke har været afgørende for den endelige vurdering af opgaven, men derimod den subjektive norm.

Med dette in mente kan hypotesen H2 bekræftes. Den kvantitative analyse har påvist, at andelen af gymnasielærerne, der er uenige i feedbacken genereret af AI, er større i AI bias-gruppen end i eksperimentgruppen *Second Opinion Influence*. Da alt andet i surveyeksperimentet har været holdt konstant med undtagelse af beslutningsstøtteværktøjets ophav, må denne forskel i tilslutning skyldes, at det er kunstig intelligens, der har genereret feedbacken. Derudover er det påvist, at de gymnasielærere i AI bias-gruppen, der ikke kan tilslutte sig feedbacken genereret af AI, har et markant højere samlet fejlniveau end de lærere, der kan tilslutte sig feedbacken. På baggrund af teorien TAM og undersøgelsens kvalitative bidrag fra lærernes egen feedback, er det sandsynliggjort, at flere faktorer fra TAM spiller en central rolle i denne tilslutning. Outputtets kvalitet, den subjektive norm, lærernes image og villighed påvirker den opfattede brugbarhed af beslutningsstøtteværktøjet, hvilket i sidste ende fører til mere eller mindre anvendelse. Dette konkluderende resultat er dermed en konvergent validering, hvormed undersøgelsens resultater er stærkere underbygget ved triangulering.

6. Konklusion

På baggrund af analysen ønskes det at besvare projektets problemformulering:

I hvilken udstrækning vil brugen af AI som beslutningsstøtteværktøj højne korrektheden i feedback- og karaktergivning i skriftlig dansk på A-niveau i gymnasiet?

Brugen af AI som beslutningsstøtteværktøj vil højne korrektheden af karaktergivningen i skriftlig dansk på A-niveau i gymnasiet. Denne undersøgelses resultater bevidner en tydelig reducering af det samlede fejlniveau i forbindelse med karaktergivningen, når beslutningsstøtteværktøjet anvendes. Denne reducering i fejlniveau stammer primært fra reducering af bias, mens støjen ikke er reduceret nævneværdigt.

Analysens hypotese H1a: *Gymnasielærere, som har benyttet AI som beslutningsstøtteværktøj, vil i gennemsnit lave færre fejl i karaktergivningen i forhold til kontrolgruppen*, kan bekræftes.

Det fremgår af analysen, at anvendelsen af beslutningstøtteværktøjet højner korrektheden af bedømmelser af danskopgaver i gymnasiet. Ved brug af *nudging*, som er en af metoderne inden for *debiasing,* er det lykkedes at mindske det samlede fejlniveau for gruppen AI-bias med 26,3 procent målt op imod kontrolgruppen. Ud af det samlede fejlniveau var det dog primært gymnasielærernes bias, der blev reduceret ved eksperimentets intervention, hvilket ikke er overraskende, da *nudging* og *debiasing* netop påvirker respondenternes bias i en given retning ved udnyttelse af tesen om konklusionsbias og den *social influence*, der sker ved en *second opinion*. Der forekommer en reducering af bias på henholdsvis 73,3 procent for SOI-gruppen og 85 procent for AI-bias-gruppen, hvilket betyder, at lærernes bedømmelsespraksis er påvirket markant i forhold til normalen. Da lærernes bias har flyttet sig tættere på den sande karakter betyder det, at anvendelsen af beslutningsstøtteværktøjet har gjort den gennemsnitlige bedømmelse mere korrekt.

Omvendt viste det sig, at interventionen ikke havde nogen nævneværdig påvirkning på det gennemsnitlige støjniveau for brugerne af beslutningsstøtteværktøjet. Med andre ord ændrede gymnasielærernes spredning og dermed præcision i karaktergivningen sig ikke, hvilket ikke er underligt, da interventionen ikke direkte har haft til formål at reducere støjniveauet.

Anvendelsen af beslutningsstøtteværktøjet har derfor, set i det store billede, i gennemsnit gjort karaktergivningen mere korrekt, men til gengæld ikke mere præcis.

Analysens hypotese H1b: *Gymnasielærere, der er eller har været censorer og som anvender beslutningsstøtteværktøjet, vil have et lavere samlet fejlniveau i forhold til de lærere, der ikke har censorerfaring*, kan ikke bekræftes. Censorerne har omtrent det samme samlede fejlniveau på tværs af alle tre grupper. Til gengæld har de lærere, der ikke har censorerfaring, et højt samlet fejlniveau i forhold til censorerne i kontrolgruppen. Dette betyder, at *boosting* højner korrektheden af bedømmelserne. Ses der på eksperimentgrupperne, findes der dog ikke nogen nævneværdig forskel i det samlede fejlniveau i forhold til om læreren har censorerfaring eller ej. Dette resultat er til gengæld et uventet fund. Undersøgelsen viser, at beslutningsstøtteværktøjet understøtter lærere uden censorerfaring så meget, at deres samlede fejlniveau bliver nærmest identisk med dem, der har censorerfaring. Ved at sammenligne forskellen mellem lærere med og uden censorerfaring i kontrolgruppen med lærere med og uden censorerfaring i eksperimentgrupperne bliver dette tydeligt. I kontrolgruppen er det samlede fejlniveau 44,7 procent mindre for de lærere, der har erfaring som censor i forhold til dem, der ikke har censorerfaring. Hvis beslutningsstøtteværktøjet ikke havde differentieret effekt på lærere afhængigt af censorstatus, kunne den samme forskel forventes i eksperimentgrupperne. Forskellen er til gengæld udlignet, hvilket må betyde, at *boosting* i forbindelse med karaktergivning bliver mindre vigtigt, når der gøres brug af beslutningsstøtteværktøjet.

Til gengæld viste undersøgelsen endvidere, at gymnasielærere i SOI-gruppen, der har erfaring som censor, har 95,3 procent mindre bias end dem, der ikke har været censor. Dette er dog ikke signifikant inden for undersøgelsens 90 procent konfidensinterval. Lærerne, der ikke har censorerfarings bias, er meget positive og dette resulterer i, at disse lærere, efter anvendelsen af beslutningsstøtteværktøjet, bedømmer generøst i forhold til lærerne, der ved, at feedback og karakter er genereret af AI.

Der viser sig ydermere en skepsis over for kunstig intelligens mellem de to grupper, der modtager intervention. Gruppen af censorer, der bliver fortalt, at den feedback de skal forholde sig til er skabt af AI, har 325 procent højere bias end SOI-gruppen.

Analysens hypotese H2: *Eksperimentgruppen AI bias vil være mere tilbøjelig til at være uenige med feedback- og karaktergivning genereret af AI i forhold til eksperimentgruppen Second Opinion Influence*, kan bekræftes. Netop her kommer den førnævnte skepsis til syne. 61 procent af gruppen AI-bias kan ikke tilslutte sig den angivne feedback. Dette er primært grundet kvaliteten af det output, som sprogmodellen har genereret. På baggrund af kvalitativ

analyse forekommer det desuden, at kvaliteten af modellens output er stærkt forbundet med respondenternes opfattelse af relevans for deres job. Dog viser den kvalitative analysen, at flere af de respondenter, der har tilkendegivet, at de ikke opfatter feedbacken som brugbar, faktisk er enige med den; men blot afskriver sig at anvende den grundet deres brug af system-1-tænkning og deraf associering med snyd og plagiat, når det gælder sprogmodeller i det danske skolesystem. Samtidig kan det også konkluderes, at de af respondenterne, der opfatter feedbacken som brugbar, og dermed anses som at have accepteret den, rykker sig meget tættere på den sande karakter. Mellem de to grupper, der modtager interventionen, er deres samlede fejlniveau henholdsvis 241 og 266,4 procent højere, når de ikke kan tilslutte sig feedbacken, som er genereret af kunstig intelligens. Dermed har den opfattede brugbarhed stor indvirkning på, hvilken effekt beslutningsstøtteværktøjet har for gymnasielæreren i sidste ende. Hvis gymnasielæreren ikke mener, at det er brugbart, vil denne fortsætte med at lave fejl, være biased og støje. Hvis de derimod kan tilslutte sig beslutningsstøtteværktøjet, så falder det samlede fejlniveau i karaktergivningen markant.

7. Diskussion

På baggrund af analysens resultater kan det bevises, at AI som beslutningsstøtteværktøj er et effektivt redskab til at højne korrektheden i lærernes karaktergivningsproces. Det må dog samtidig også anerkendes, at lærernes tydelige skepsis og modstand af generativ kunstig intelligens er så omfattende, at flere af lærerne derfor igennem system-1-tænkning frasiger sig dens feedback, hvor de egentlig er enige med den. Derfor findes det relevant at diskutere årsagen til dette.

Som tidligere nævnt er Gemini, som de fleste andre sprogmodeller, baseret på *'deep learning'* og besidder derfor et så omfattende usynligt neuralt netværk af data, at det for den enkelte person er umuligt at kortlægge, hvordan dens output er dannet. Som beskrevet i afsnit 4.4.4 er det forsøgt at undgå for meget inkorporering af eksterne data i projektets genererede feedback på baggrund af det anvendte prompts udformning. Det må derfor overvejes, om netop denne eksterne data kan være årsagen til gymnasielærernes skepsis. Heraf kan det diskuteres, hvordan en sprogmodel, der skal anvendes som beslutningsstøtteværktøj, skal opbygges, så det usynlige lag undgås.

Projektets prompt er udformet på baggrund af læringsmålene for dansk på A-niveau for det almene gymnasium. Hvis man programmerer en sprogmodel således, at det neurale netværks usynlige lag udelukkende består af krav, regler og mål for enten hele faget eller den enkelte opgave, vil mistroen til det endelige output kunne reduceres. Denne model kunne eventuelt udarbejdes af Børne- og Undervisningsministeriet eventuelt i samarbejde med dansklærere, fagkonsulenter eller andre relevante interessenter for på den måde at blive blåstemplet. En sprogmodel, der er udarbejdet i samarbejde med og på baggrund af inputs fra fagpersoner, kan derfor forventes at være mere pålidelig. Deraf må det forventes, at de eksterne faktorer i TAM ikke vil have den samme indvirkning som i denne undersøgelse, da teknologien vil være mere alment accepteret, og gymnasielærerens opfattelse af dens relevans for deres job vil være større. Således vil flere gymnasielærere opfatte teknologien som brugbar, acceptere den og i sidste ende også anvende den.

Tanken om den 'perfekte' bedømmer i form af en gennemarbejdet akademisk og fagligt baseret sprogmodel leder op til videre tanker om mulighederne heraf. Som eksempel kan den nuværende praksis med to bedømmere på de skriftlige afgangseksamener ved de danske gymnasier omstruktureres således, at der kun vil være en bedømmer i samarbejde med beslutningstøtteværktøjet baseret på AI. Dette vil da fjerne de både tids- og udgiftsmæssige

problematikker, som der er ved at have to lærere til en opgavebesvarelse. Det beløb, der frisættes ved at fjerne en bedømmer, vil kunne anvendes et andet sted på uddannelsesområdet. Hertil kan nævnes eksempelvis efteruddannelse af både de nye og ældre lærere i IT-anvendelse, oplæring i generativ kunstig intelligens eller muligheden for at ansætte flere lærere, så de nutidige problematikker, der er nævnt i projektet, kan afhjælpes. Dette åbner potentielt det danske skolesystem op og skaber friere rammer for den enkelte lærer, hvilket gør det muligt for dem at nå deres egentlige kerneopgave.

Det kan dog diskuteres, hvorvidt et beslutningsstøtteværktøj baseret på kunstig intelligens, vil være til gavn for eleverne på de danske gymnasier. Som det er beskrevet i projektet, er kunstig intelligens støjløs, og kan derfor give den mest objektive vurdering af en given opgavebesvarelse. Dette er ud fra et fagligt perspektiv påskønnet, da man derved vil komme så tæt på den sande og derfor "rigtige" karakter. Der er dog en ting, som AI ikke kan, og formentlig aldrig vil komme til at kunne - nemlig at føle. Denne undersøgelse fokuserer på gymnasielærere og primært den faglige del af deres arbejde, og har dermed også fokus på den summative vurdering, de foretager. Som nævnt foretager de også formative vurderinger, når de skal give karakter til deres egne elever, som følger deres fag. Denne disciplin kræver et medfølende samt personligt forhold til eleven, da den skal kunne forholde sig specifikt til den enkelte elevs faglige dannelse og udvikling gennem gymnasiet. Det kan en robot ikke. Det skal også derfor understreges, at denne undersøgelse ikke ønsker at fjerne lærerne helt fra ligningen. I stedet bør introduktionen af AI anses som et værktøj, der netop støtter de fagprofessionelle i at skønne i forbindelse med karaktergivning- og beslutningsprocessen. For det er netop her, at filmen knækker for AI og generative sprogmodeller.

Gemini er den første sprogmodel til at passere menneskets evne til at multitaske og løse problemstillinger i den tidligere nævnte MMLU-benchmarking test. Derfor kan den i sin "perfekte" udformning, uden indvirkning fra det usynlige lag, anses som overlegen målt mod en gymnasielærer i dansk, når det gælder vurderinger af opgaver. Men den kan ikke skabe relationer, vise omsorg eller aflæse mimikken på en elev på samme måde, som en lærer kan. Samtidig kan den ikke stå for dannelsen af de studerende på Danmarks gymnasier. Beslutningsstøtteværktøjet vil til gengæld forhåbentligt frigøre tid til læreren, så der netop bliver tid til at sørge for, at færre elever bliver tabt på vejen. Det er derfor sandsynligt, at en fornuftig og faglig programmering, samt en god implementering af AI som en ny kollega på gymnasiet, har potentialet til at styrke det danske gymnasie og frigøre mere tid til gymnasielærernes kerneopgave.

8. Litteraturliste

Agresti, A. (2018). *Statistical Methods for the Social Sciences* (5. ed.). Global Edition.

Aarhus Universitet. (2024). *Metodeguiden*. Retrieved from Aarhus Universitet: https://metodeguiden.au.dk/eksperimenter

Cabitza, F., Campagner, A., Angius, R., Natali, C., & Reverberi, C. (2023, april 18). AI Shall Have No Dominion: on How to Measure Technology Dominance in AI-supported Human decision-making. *Conference on Human Factors in Computing Systems*.

Chuttur, M. (2009, september 27). *Overview of the Technology Acceptance Model: Origins, Developments and Future Directions.* Retrieved from All Sprouts Content: https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1289&context=sprouts_all

Clement, S. L. (2017). *Survey: Design, Stikprøve, Spørgeskema & Analyse* (1. ed.). Hans Reitzels Forlag.

Clement, S. L., & Ingemann, J. H. (2019). *Introduktion til praktisk statistik* (3. ed.). Syddansk Universitetsforlag.

Dansk Sprognævn. (2023, december 8). *ChatGPT er kåret til årets ord 2023*. Retrieved april 1, 2024, from Dansk Sprognævn: https://dsn.dk/nyheder-og-arrangementer/chatgpt-er-kaaret-til-aarets-ord-2023/

Danske Gymnasier. (2024). *Find Gymnasier*. Retrieved from Danske Gymnasier: https://danskegymnasier.dk/find-gymnasier/

Danske Patienter. (2024). *Beslutningsstøtteværktøj – til alle patientgrupper*. Retrieved from Danske Patienter: https://danskepatienter.dk/beslutningsstoettevaerktoej-til-alle-patientgrupper

EVA. (2016). *Karaktergivning i gymnasiet.* Retrieved from Danmarks Evalueringsinstitut: https://eva.dk/Media/638379657332597157/Karaktergivning%20i%20gymnasiet.pdf

EVA. (2017). *Kortlægning og analyse af karaktergivningen – baggrundsrapport til evaluering af 7- trins-skalaen 2007-2016.* Retrieved from Danmarks Evalueringsinstitut: https://www.uvm.dk/-/media/filer/uvm/adm/2018/pdf18/jun/180613-evaluering-af-7-trins-skalaen-fase-1.pdf

EVA. (2020). *Forsøg med karakterfri 1.g. Slutrapport for skoleårene 2017/2018 og 2018/2019.* Retrieved from eva.dk: https://eva.dk/Media/638372874845424505/Forsøg%20med%20karakterfri%201.g%20-%20slutrapport.pdf

Finansministeriet. (2022, maj). *Danmarks digitaliseringsstrategi.* Retrieved from https://www.regeringen.dk/media/11324/danmarks-digitaliseringsstrategi-sammen-om-den-digitale-udvikling.pdf

Floridi, L., & Chiriatti, M. (2020, december). GPT-3: Its Nature, Scope, Limits, and Conse quences. *Minds and Machines*, pp. 1-14.

Frederiksen, M. (2020). Kapitel 11: Mixed methods-forskning. In S. Brinkmann, & L. Tanggaard, *Kvalitative metoder - en grundbog* (3. ed.). Hans Reitzels forlag.

Google DeepMind. (2024). *Gemini Models*. Retrieved from Google DeepMind: https://deepmind.google/technologies/gemini/#gemini-1.0

Gymnasieskolernes Lærerforening. (2019). *Har besparelserne på gymnasieområdet konsekvenser for eleverne?* Gymnasieskolernes Lærerforening.

Hong, X., Zhang, M., & Liu, Q. (2021, juni). Preschool Teachers' Technology Acceptance During the COVID-19: An Adapted Technology Acceptance Model. *Frontiers in psychology*(12), pp. 1-11.

IBM. (2024). *What is artificial intelligence (AI)?* Retrieved april 1, 2024, from IBM: https://www.ibm.com/topics/artificial-intelligence

Ingemann, J. H. (2020). *Videnskabsteori for økonomi, politik og forvaltning* (4. ed.). Samfundslitteratur.

Jacobsen, A. V., Erenbjerg, A., Törnfelt, C., & Tassy, A. (2023, december). *It-anvendelse i befolkningen.* Retrieved from Danmarks Statistik: https://www.dst.dk/Site/Dst/Udgivelser/GetPubFile.aspx?id=49775&sid=itbef2023

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A Flaw in Human Judgement.* London: William Collins.

Kjeldsen, K. N., & Larsen, K. B. (2021). *Karakterforskelle på de gymnasiale uddannelser.* CEPOS.

Krause-Jensen, J., Kamp, A., Nielsen, M. L., & Spanger, M. (2022). Arbejdsliv 4.0: digitalisering og kunstig intelligens i arbejdet. *Tidsskrift for Arbejdsliv*, pp. 5-9.

Lindebjerg, J., & Rahr, H. B. (2017, oktober 17). Lille cancer – stort dilemma. *Ugeskr Læger*, p. 1921.

Nielsen, L. T. (2020, april 27). *Elever i udskolingen ville ønske, at de gik mindre op i karakterer*. Retrieved from Danmarks Evalueringsinstitut: https://eva.dk/debat/2020/apr/elever-i-udskolingen-ville-oenske-at-de-gik-mindre-op-i-karakterer

Roersen, A. B., Nielsen, J. F., Andreasen, A.-S. B., Laursen, M., Svendsen, E., Jansson, M. B., & Sachse, J. W. (2022). *Tilfældighedernes karaktergivning - et studie om bedømmelsespraksissens betydning for karakterernes troværdighed.* Aalborg: Aalborg Universitet, Det samfundsvidenskabelige fakultet.

Romme-Mølby, M. (2023, juni 17). *Lærere og elever uenige om ChatGPT i undervisningen*. Retrieved april 1, 2024, from Gymnasieskolen: https://gymnasieskolen.dk/articles/laerere-og-elever-uenige-om-chatgpt-i-undervisningen/

Styrelsen for Undervisning og Kvalitet. (2024, januar 8). *Bliv skriftlig censor og styrk din faglighed.* Retrieved april 1, 2024, from Børne- og undervisningsministeriet: https://www.uvm.dk/-/media/filer/uvm/udd/gym/pdf24/jan/240108bliv-skriftlig-censor-og-styrk-din-faglighed.pdf

UVM. (2020, september). *Modeller for en ændret karakterskala.* Retrieved from Børne- og Undervisningsministeriet: https://www.uvm.dk/-/media/filer/uvm/aktuelt/pdf20/sep/200930-modeller-for-en-aendret-karakterskala-september-2020.pdf

UVM. (2021). *Dansk A, stx, vejledning.* Retrieved from Børne- og Undervisningsministeriet: https://www.uvm.dk/-/media/filer/uvm/gym-vejledninger-til-laereplaner/03082020/210816-dansk-a-stx-vejledning-juni-2021.pdf

UVM. (2023c, august 2). *Anvendelse af 7-trins-skalaen*. Retrieved april 1, 2024, from Børne- og Undervisningsministeriet: https://www.uvm.dk/uddannelsessystemet/7-trins-skalaen/anvendelse-af-7-trins-skalaen

UVM. (2024b, april 15). *Censormøde 2024*. Retrieved from Børne- og Undervisningsministeriet: https://www.uvm.dk/gymnasiale-uddannelser/proever-og-eksamen/om-censur/censormoede

UVM. (2023b, august 2). *Karakterer på 7-trins-skalaen*. Retrieved from Børne- og Undervisningsministeriet: https://www.uvm.dk/uddannelsessystemet/7-trins-skalaen/karakterer-paa-7-trins-skalaen

UVM. (2023d, januar 23). *Normer for skriftlig censur ved de gymnasiale uddannelser.* Retrieved from Styrelsen for Undervisning og Kvalitet: https://www.uvm.dk/-/media/filer/uvm/udd/gym/pdf23/jan/230125-normer-for-skriftlig-censur-ved-de-gymnasiale-uddannelser.pdf

UVM. (2023a, august 11). *Standpunktskarakterer*. Retrieved april 1, 2024, from Børne- og undervisningsministeriet: https://www.uvm.dk/folkeskolen/folkeskolens-proever/faglig-forberedelse/standpunktskarakterer

UVM. (2024a, maj 6). *Skriftlige censorer*. Retrieved maj 6, 2024, from Børne- og undervisningsministeriet: https://www.uvm.dk/gymnasiale-uddannelser/proever-og-eksamen/om-censur/skriftlige-censorer

Venkatesh, V., & Davis, F. D. (2000, februar 1). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science*(46), pp. 186-204.