Kunstig forvaltning. En empirisk test af store sprogmodellers kapacitet til at foretage forvaltningsafgørelser

Temanummer: Ret og rimelighed i kommunal sagsbehandling

Kan ChatGPT 4 producere afgørelser, der er tilstrækkelige som forvaltningsafgørelser hos Jobcenter Vejen? Med udgangspunkt i et katalog af konkrete afgørelser fra Jobcenter Vejen og et katalog af det bagvedliggende sagsmateriale tester vi ved hjælp af et dommerpanel, der bedømmer afgørelserne blindt, hvorvidt ChatGPT 4 kan producere tilstrækkelige afgørelser på baggrund af det foreliggende sagsmateriale. Vi finder overordnet, at der ikke er en statistisk signifikant forskel i dommerpanelets bedømmelse af henholdsvis afgørelser produceret af ChatGPT 4 og afgørelser produceret af Jobcenter Vejen. Vi konkluderer derfor, at ChatGPT 4 kan producere tilstrækkelige afgørelser. Vi bemærker dog, at denne konklusion er begrænset til vores datasæt, og at yderligere undersøgelser er påkrævet for at kunne drage en mere entydig konklusion. På baggrund af denne overordnede konklusion diskuterer vi kort forskellige implikationer af ChatGPT 4's kapacitet til at producere tilstrækkelige forvaltningsafgørelser, herunder etiske og juridiske implikationer.

JAKOB GRAULUND JØRGENSEN,
Adjunkt, Institut for samfund og kommunikation, UC SYD.
jgjo@ucsyd.dk.

LOUISE HANSEN
Lektor, Institut for samfund og kommunikation, UC SYD.
lhan@ucsyd.dk.

Kan ChatGPT træffe tilstrækkelige forvaltningsafgørelser hos Jobcenter Vejen?

ChatGPT har allerede indfundet sig i retslokaler over hele verden (Smith o.a., 2023; Staff, 2023; Taylor, 2023). Men brugen af ChatGPT og andre store sprogmodeller til at træffe juridiske afgørelser er ingenlunde uproblematisk. Brugen af store sprogmodeller til at træffe beslutninger med retlig virkning er således behæftet med både etiske, juridiske og praktiske problemstillinger. Helt grundlæggende kan tre spørgsmål opstilles: er det rimeligt, er det lovligt og er det muligt.

I denne artikel tager vi hul på det sidste af disse tre spørgsmål. Dvs. er det muligt for store sprogmodeller som f.eks. ChatGPT at træffe tilstrækkelige afgørelser. Hvis konklusionen på dette spørgsmål er negativ, er de første to spørgsmål alene af teoretisk interesse. Omvendt vil en positiv konklusion gøre begge spørgsmål praktisk relevante.

» Vores dommerpanel bedømmer således i et blindt format både de faktiske afgørelsesbreve produceret af Jobcenter Vejen og afgørelsesbreve produceret af ChatGPT 4

Konkret tester vi, om ChatGPT 4 (GPT-4) kan producere afgørelser, der er tilstrækkelige som forvaltningsafgørelser hos Jobcenter Vejen. Dette gør vi vha. sagsmateriale fra Jobcenter Vejen og et dommerpanel, der i en blindtest bedømmer kvaliteten af ChatGPT's afgørelser. Vi har med hjælp fra medarbejdere hos Jobcenter Vejen udvalgt 26 forskellige afgørelser, der spænder over et bredt udsnit af jobcenterets portefølje af opgaver. Disse 26 afgørelser har hos Jobcenter Vejen afstedkommet 30 afgørelsesbreve. Det er disse afgørelsesbreve, vi tager som udgangspunkt for vores test. Vores dommerpanel bedømmer således i et blindt format både de faktiske afgørelsesbreve produceret af Jobcenter Vejen og afgørelsesbreve produceret af ChatGPT 4. Vores er mål er ikke at drage en endegyldig konklusion om store sprogmodellers kapacitet til at foretage forvaltningsafgørelser, men alene at afsøge, om det overhovedet er relevant at undersøge yderligere. Vi har således en eksplorativ tilgang, der ønsker at afsøge om store sprogmodellers er relevante ift. den forsatte effektivisering af den offentlig sektors afgørelsesvirksomhed.

Vi har udvalgt Jobcenter Vejen som udgangspunkt for vores test, idet jobcenteret producerer en løbende strøm af forskelligartede afgørelser med tilhørende baggrundsmateriale, hvorfra en passende stikprøve kan udtrækkes. Samtidig har beskæftigelsesområdet historisk været underlagt et betydeligt effektiviseringspres. Det er således et område af den offentlige sektor, hvor det kan forventes, at nye automatiseringsværktøjer som kunstig intelligens hurtigt vil blive taget i brug, hvis de viser sig praktisk anvendelige.

Store sprogmodellers opgaveagnostiske potentiale

Store sprogmodeller udspringer af et arbejde startet i 1940'erne med det formål at gøre computere i stand til at forstå og producere almindelig tale (natural language) (Jiang og Lu, 2020: 211). Feltet delte sig tidligt i to grene. Den ene gren forfulgte en regelbaseret semantisk tilgang (symbolist), mens den anden gren forfulgte en frekventistisk tilgang (frequentist) (Jiang og Lu, 2020: 211). Store sprogmodeller er en udløber af den frekventistiske tilgang, der med øget computerkraft og nye algoritmiske tilgange, herunder særligt neurale netværk (ibid.: 212) har vist sig som (foreløbigt) det mest produktive spor.

Særligt fremkomsten af neurale netværk er central for udviklingen af store sprogmodeller og denne tilgang er i overvejende grad opgaveagnostisk (Brown o.a., 2020: 1). Dvs. modellerne trænes ikke til at løse én specifik opgave (f.eks. oversættelse, tekstklassificering eller lignende). I stedet er arkitekturen og fortræningen (pretraining) opgaveuafhængig (ibid.). Efter endt træning kan den enkelte model underlægges træning, der fintuner modellen ift. en eller flere specifikke opgaver (ibid.).

Den opgaveagnostiske tilgang gør, at store sprogmodeller har potentiale til at kunne løse opgaver, der falder udenfor den opgavekreds, de oprindeligt var tiltænkt. Herved placerer store sprogmodeller sig i et spændingsfelt mellem

specialiseret kunstig intelligens og almen kunstig intelligens, hvorved forstås intelligens, der er overførbar mellem forskelligartede domæner (Bubeck o.a., 2023). Juridiske analyser er netop en af de opgaver, der falder udenfor den opgavekreds, store sprogmodeller oprindeligt var tiltænkt.

Forskningsstatus

Selvom avancerede sprogmodeller først for alvor trådte ud i rampelyset med ChatGPT i november 2022, er der allerede foretaget adskillige studier af både store sprogmodellers kapacitet mht. juridisk analyse og ift. forudsigelser – ved forudsigelser forstår vi både fremskrivninger og slutninger på baggrund af datasæt, og det er dette sidste element af forudsigelser, der er relevant ift. afgørelser, idet vi netop ønsker at undersøge, om sprogmodellen kan forudsige den korrekte afgørelse på baggrund af relevant sagsmateriale.

OpenAI (2023) viser således i relation til deres egen model GPT-4, at denne model er i stand til at gennemføre LSAT (The Law School Admission Test) på et niveau svarende til den 88. percentil (ibid.: 5) og bestå en simuleret version af the Uniform Bar Examination – en standardiseret juridisk eksamen, som anvendes til at vurdere kvalifikationerne hos personer, der ønsker at blive advokater (NCBE, u.å.) – med en præstation blandt de bedste 10 pct. (OpenAI, 2023: 6).

Ift. store sprogmodeller og forudsigelser viser Lopez-Lira og Tang (2023), at avancerede sprogmodeller i modsætning til tidligere iterationer kan bruges til at forudsige bevægelser på aktiemarkedet baseret på sprogmodellens tolkning af nyhedsoverskrifter. Liang o.a. (2023) viser, at en sprogmodelsramme kan bruges til at forudsige gruppers bevægelsesmønstre ved offentlige begivenheder. Kang o.a. (2023) viser, at sprogmodeller kan bruges til at forudsige brugerpræferencer på baggrund af tidligere adfærd. Ekambaram o.a. (2024) viser, at sprogmodelsarkitektur kan bruges til tidsseriefremskrivning. Faria-E-Castro og Leibovici (2023) viser, at sprogmodeller kan bruges til at forudsige inflationsniveauet.

Andre studier af store sprogmodeller viser dog, at modellerne også har betydelige problemer med nogle typer af forudsigelser. Schönegger og Park (2023) viser således, at GPT-4 underpræsterer ift. både mennesker og tilfældige gæt, når modellen skal foretage forudsigelser om fremtidige begivenheder. Park o.a. (2024) viser, at store sprogmodeller i visse sammenhænge har en tendens til at mangle diversitet i deres svar. Dvs. sprogmodellerne søger at give et "korrekt" svar uden at tage højde for en relevant kontekst.

Samlet er forskningssituationen således uafklaret ift. store sprogmodellers kapacitet til at forudsige indenfor en lang række domæner. Store sprogmodellers formåen ift. juridiske analyser i en dansk kontekst er også uafklaret.

Kunstig intelligens i dansk forvaltning

Kunstig intelligens er allerede i dag i brug i den danske offentlige forvaltning (Loiborg, 2020: 64). Også generativ kunstig intelligens, som store sprogmodeller er en delmængde af, er allerede i aktiv brug (KL og Digitaliseringsstyrelsen, 2024).

Det må således kunne antages, at det er muligt at anvende store sprogmodeller i den faktiske forvaltningsvirksomhed – store sprogmodeller er således allerede blevet indført[1]. En væsentlig del af den offentlige forvaltning er dog ikke blot beskæftiget med faktisk forvaltningsvirksomhed, men også afgørelsesvirksomhed. For indeværende kan det antages, at brugen af store sprogmodeller og formentlig al kunstig intelligens baseret på black box maskinlæring ikke kan bruges til afgørelsesvirksomhed uden at overtræde almindelige forvaltningsretlige principper (Loiborg, 2020: 72), medmindre lovgiver har givet lovhjemmel til at fravige disse (ibid.).

Vi finder det derfor rimeligt at antage, at store sprogmodeller ikke allerede (systematisk) benyttes til foretage forvaltningsafgørelse. Vi ønsker med vores undersøgelse at undersøge, hvorvidt store sprogmodeller kunne bruges til afgørelser. Vi sætter således parentes om lovlighedsspørgsmålet og de etiske spørgsmål og hopper direkte videre til det praktiske spørgsmål om faktisk anvendelse.

En empirisk test hos Jobcenter Vejen

Udformning af eksperimentet

For at gennemføre en test af store sprogmodellers kapacitet til at foretage forvaltningsafgørelser har vi i samarbejde med Jobcenter Vejen udarbejdet et katalog af allerede trufne afgørelser og det bagvedliggende sagsmateriale til hver enkelt afgørelse. Vi har i alt undersøgt 26 afgørelser. Disse sager er alle blevet anonymiseret af Jobcenter Vejen, og vi har ikke haft adgang til de personfølsomme elementer af sagerne – dvs. navne, adresser eller andet, der ville kunne lede til en entydige identifikation af en person. Disse oplysninger har dog – i vores vurdering – ikke relevans for afgørelserne, der træffes i de 26 sager. De 26 sager har i alt affødt 30 breve til borgere eller virksomheder, som omhandler trufne afgørelser – f.eks. en skriftlig afgørelse om afslag på sygedagpenge. De 26 afgørelser spænder over ansøgninger om sygedagpenge, refusionssager, ansøgninger om uddannelseshjælp, sager om ressourceforløbsydelser, sager om bevilling af brøkpension, sager om tilskud til voksenlærling, ansøgninger om f.eks. kursusforløb og generhvervelse af stort kørekort, og sager om transportudgifter. Sagsmaterialet til de forskellige sager spænder derfor over et bredt spektrum. Nogle sager har nogle få bilag i form af dokumentation for rejseomkostninger, mens andre sager dækker over et længere forløb med journalnotater, skrivelser mv. mellem jobcenter og borger.

For at gennemføre en test af store sprogmodellers kapacitet til at foretage forvaltningsafgørelser har vi i samarbejde med Jobcenter Vejen udarbejdet et katalog af allerede trufne afgørelser og det bagvedliggende sagsmateriale til hver enkelt afgørelse

Som led i vores eksperiment har vi fjernet de endelige afgørelser – dvs. udsendte breve – fra sagsmaterialet, hvorved vi har fået et katalog af sagsmateriale uden afgørelse og et katalog af faktisk trufne afgørelser. Herefter har vi vha. OCR digitaliseret det modtagne sagsmateriale med henblik på at indlæse det som kontekst i en sprogmodel.

Som vores store sprogmodel har vi valgt at benytte ChatGPT 4 (GPT-4). GPT-4 var på forsøgstidspunkt den førende model i markedet og er (medio 2024) fortsat blandt de ledende modeller (LMSYS Chatbot Arena Leaderboard, 2024). GPT-4 er samtidig i den familie af modeller mange administrative medarbejdere har adgang til i kraft af Microsoft Copilot (Mehdi, 2024). Vi har som led i eksperimentet også testet en enkelt sag (M2 – S2) med Claude Opus 3 (Anthropic, 2024).

For at undgå uforvarende at hjælpe GPT-4 ved udformningen af de enkelte afgørelser har vi udviklet et fælles instruktionssæt (prompt) som vi giver GPT-4. Vi benytter i den forbindelse OpenAI's mulighed for at lave såkaldte GPTs. Vi kalder vores GPT for GPT-sagsbehandler

I udførelsen af eksperimentet giver vi GPT-sagsbehandler sagsmaterialet for hver enkelt sag uden yderligere instruks – dvs. GPT-4 har kun én enkelt sag i kontekst ad gangen. GPT-sagsbehandler svarer på det oploadede materiale med et afgørelsesbrev, jf. det fælles instruktionssæt. Til hver ny sag starter vi en ny samtale med GPT-sagsbehandler for at sikre et rent kontekstvindue – vi har i den forbindelse bevidst slået ChatGPT's hukommelsesfunktion fra. Vi benytter i vores eksperiment altid første svar, der afgives. I tre sager (M1 – S1A, M1 – S1B og M1-S5) giver vi følgende yderligere instruktion: "Skriv også et brev til arbejdsgiveren." GPT-sagsbehandler udformer på den baggrund et yderligere afgørelsesbrev. I vores test af Claude Opus 3 benytter vi samme instruktionssæt som til GPT-4. I en enkelt sag (M3 – S6) benytter vi følgende yderligere instruktion: "Gennemgå sagen igen. Du har allerede alle de dokumenter du efterspørger. Lav et nyt fuldt svar." GPT-sagsbehandler udformer på den baggrund et nyt afgørelsesbrev. Til sag M3 – S6 har dommerne både fået første svar fra GPT-sagsbehandler og det reviderede svar på baggrund af den yderligere instruks. Til M2 – S2 har dommerne både fået svar fra GPT-sagsbehandler og Claude Opus 3. Til M1 – S1A, M1 – S1B og M1-S5 har dommerne fået både breve til borger og til virksomhed fra både GPT-sagsbehandler og fra jobcentret.

Sammensætning af dommerpanel

Til at bedømme de fremstillede afgørelser har vi samlet et dommerpanel på tre dommere. Dommerne har modtaget de i alt 62 afgørelse – 30 faktiske afgørelse, 30 sager fra GPT-sagsbehandler, 1 revideret sag fra GPT-sagsbehandler og 1 sag fra Claude Opus 3 – i et blindt format, hvor det ikke er markeret, hvorvidt afgørelsen er en faktisk afgørelse, eller om den er produceret af en stor sprogmodel. I den forbindelse har vi også digitaliseret de faktiske afgørelser vha. OCR og overført dem til et word-format uden brevhoved. Herved undgår vi, at det er umiddelbart klart, hvilke sager der er faktiske afgørelser. Vi har foruden disse ændringer i formatering også fjernet navnene på sagsbehandlerne i de faktiske afgørelser og erstattet med følgende "[navn_sagsbehandler]" eller et lignende format f.eks. "[navn_ydelsesrådgiver]". Dette er igen gjort for at undgå, at det er umiddelbart klart, hvilke sager der er faktiske afgørelser.

Dommerpanelet består af en leder af fra Jobcenter Vejen med mere end 25 års erfaring på arbejdsmarkedsområdet, en tidligere direktør for Arbejdsmarkedsområdet, herunder et jobcenter i en mellemstor danske kommune, og en juridisk ph.d. med tidligere ansættelse i et jobcenter i en større dansk kommune og i Ankestyrelsen med erfaring i behandling af klager over afgørelser. Dommerpanelet er udvalgt på baggrund af deres kendskab til området. Det bemærkes i den forbindelse, at kun lederen fra Jobcenter Vejen i sin nuværende rolle fortsat er beskæftiget med området. Givet de hyppige lovændringer på området må det antages, at de to dommere uden aktuel tilknytning til området i et vist omfang besidder forældet viden. Vi vurderer dog samlet, at begge har tilstrækkelig viden til, at deres vurderinger af afgørelserne er relevante.

Dommernes bedømmelse

De tre dommere har bedømt hver enkelt afgørelse med udgangspunkt i et spørgeskema med seks spørgsmål. De fire første spørgsmål omhandler elementer af afgørelsen og bedømmes på skalaen "meget enig" (1), "enig" (2), "hverken enig eller uenig" (3), "uenig" (4), "meget uenig" (5) – kodningen fremgik ikke af spørgeskemaet. Spørgsmålene er som følger:

1. "Den skriftlige formidling er passende"
2. "De juridiske elementer er passende"
3. "Afgørelsen er korrekt"
4. "Afgørelsen er tilstrækkelig som forvaltningsafgørelse"

De to sidste spørgsmål omhandler den samlede afgørelse. Her vurderes først den samlede afgørelsen på følgende skala "meget god" (1), "god" (2), "hverken god eller dårlig" (3), "dårlig" (4), "meget dårlig" (5). Endelig vurderes, om afgørelsen er genereret ved hjælp af kunstig intelligens på følgende skala "ja" (1), "nej" (2), "ved ikke" (3).

Resultatet af vores eksperiment

I tabel 1 og 2 er udregnet den gennemsnitlige værdi for besvarelserne af afgørelsesspørgeskemaerne bedømt af hver af de tre dommere for henholdsvis faktiske sager fra Jobcenter Vejen og sager udformet af GPT-sagsbehandler. Lavere værdier angivet for alle kategorier en mere positiv bedømmelse undtaget kategorien "Vurdering om AI". Den reviderede sagsafgørelse (vedrørende M3 – S6) og Claude Opus 3's sag (vedrørende M2 – S2) indgår ikke i beregningen.

Tabel 1: Bedømmelse dommere, faktiske afgørelser

Jobcenter	Leder (n=30)	Tidl. direktør (n=30)	Juridisk ph.d. (n=30)
Skriftlig formidling	2,97	2,77	3,30
Juridiske elementer	2,80	2,43	3,10
Afgørelsen korrekt	2,10	2,27	2,47
Afgørelsen tilstrækkelig	3,10	2,30	3,83
Samlet vurdering	3,17	2,57	3,40
Vurdering om AI	1,57	1,60	2,90

Tabel 2: Bedømmelse dommere, ChatGPT-afgørelser

Kunstig intelligens	Leder (n=30)	Tidl. direktør (n=30)	Juridisk ph.d. (n=30)
Skriftlig formidling	2,17	2,93	3,07
Juridiske elementer	2,87	2,93	3,67
Afgørelsen korrekt	1,80	2,70	2,80
Afgørelsen tilstrækkelig	2,93	2,77	4,07
Samlet vurdering	2,80	3,00	3,67
Vurdering om AI	1,70	1,40	2,80

I tabel 3 er udregnet differencen mellem tabel 1 og tabel 2. En positiv difference er udtryk for, at dommeren i gennemsnit har givet GPT-sagsbehandlers afgørelser en mere positiv bedømmelse end jobcenterets afgørelser. En negativ difference er udtryk for, at dommeren i gennemsnit har givet GPT-sagsbehandlers afgørelser en mere negativ bedømmelse end jobcenterets afgørelser.

Tabel 3: Difference mellem bedømmelse af faktiske afgørelser og ChatGPT-afgørelser

Difference	Leder (n=30)	Tidl. direktør (n=30)	Juridisk ph.d. (n=30)
Skriftlig formidling	0,80	-0,17	0,23
Juridiske elementer	-0,07	-0,50	-0,57
Afgørelsen korrekt	0,30	-0,43	-0,33
Afgørelsen tilstrækkelig	0,17	-0,47	-0,23
Samlet vurdering	0,37	-0,43	-0,27
Vurdering om AI	-0,13	0,20	0,10

I tabel 4 og 5 udregnes, hvorvidt forskellen mellem de to sæt af bedømmelser er statistisk signifikant. Dvs. vi tester nulhypotesen, at der ikke er signifikant forskel i dommernes bedømmelse af henholdsvis faktiske afgørelser fra jobcenteret og GPT-sagsbehandlers afgørelser. I tabel 4 er udregnet en parret t-test, der tester hvorvidt, der er en statistisk signifikant forskel mellem middelværdierne af et par af værdier. Idet n=30, og idet bedømmelserne ikke kan antages at være normalfordelt – vi har ikke nogen grund til at antage, at dette skulle være tilfældet – foretages også Wilcoxons rangtest for parrede observationer. Fordelen ved denne test er, at den er ikke-parametrisk. Med et signifikans-niveau på 5 pct. er der ikke forskel i resultatet mellem t-testen og Wilcoxons rangtest.

Tabel 4: Parret t-test

t-test (parret)	Leder (n=30)	Tidl. direktør (n=30)	Juridisk ph.d. (n=30)
Skriftlig formidling	1,82%	53,78%	44,03%
Juridiske elementer	83,15%	6,15%	1,68%
Afgørelsen korrekt	13,03%	6,20%	20,94%
Afgørelsen tilstrækkelig	63,04%	5,04%	34,49%
Samlet vurdering	23,34%	7,92%	25,50%
Vurdering om AI	50,20%	26,38%	32,56%

*Fremhævede celler markerer spørgsmål, hvor der med over 95 pct. sikkerhed er forskel i dommernes bedømmelse af henholdsvis GPT-sagsbehandlers og Jobcenter Vejens afgørelser.

Tabel 5: Wilcoxons rangtest

Wilcoxon rangtest	Leder (n=30)	Tidl. direktør (n=30)	Juridisk ph.d. (n=30)
Skriftlig formidling	2,01%	60,71%	45,91%
Juridiske elementer	62,72%	7,93%	2,30%
Afgørelsen korrekt	16,22%	9,49%	24,61%
Afgørelsen tilstrækkelig	77,02%	7,66%	33,18%
Samlet vurdering	31,30%	10,07%	23,96%
Vurdering om AI	53,12%	32,19%	28,50%

*Fremhævede celler markerer spørgsmål, hvor der med over 95 pct. sikkerhed er forskel i dommernes bedømmelse af henholdsvis GPT-sagsbehandlers og Jobcenter Vejens afgørelser.

Vi får herved, at vi ikke med 95 pct. sikkerhed kan konkludere, at der er forskel mellem dommerens bedømmelse af afgørelser fra henholdsvis jobcenter og GPT-sagsbehandler. To dommere har dog hver ét spørgsmål, hvor vi med over 95 pct. sikkerhed kan konkludere, at der er forskel i deres bedømmelse. Vi kan her konkludere, at lederen fra jobcenteret bedømmer GPT-sagsbehandlers skriftlige formidling signifikant mere positivt end jobcenterets skriftlige formidling. Vi kan også konkludere, at den juridiske ph.d. bedømmer GPT-sagsbehandlers juridiske elementer signifikant mere negativt end jobcenterets juridiske elementer.

Nogle implikationer af GPT-4's evne til at træffe forvaltningsafgørelser

Bekræftende konklusion

Resultatet af vores eksperiment viser helt overordnet, at forskellen mellem vores dommerpanels bedømmelser af henholdsvis GPT-sagsbehandlers afgørelser og jobcenterets afgørelser ikke er betydelig nok til, at forskellen er statistik signifikant med et 95 pct. signifikansniveau. Dette er i sig selv et markant resultat. Vores GPT-sagsbehandler er ikke fintunet til eller på anden vis blevet målrettet jobcenterområdet. Alligevel kan vi generere afgørelser over 26 forskellige sager med et enkelt standardinstrukssæt, der ikke er åbenbart inferiøre i forhold til de afgørelser, medarbejderne på Jobcenter Vejen producerer.

» Resultatet af vores eksperiment viser helt overordnet, at forskellen mellem vores dommerpanels bedømmelser af henholdsvis GPT-sagsbehandlers afgørelser og jobcenterets afgørelser ikke er betydelig nok til, at forskellen er statistik signifikant med et 95 pct. signifikansniveau

Vi kan således konkludere bekræftende i forhold til vores undersøgelsesspørgsmål. Dvs. vi kan konkludere, at GPT-4 kan producere afgørelser, der er tilstrækkelige som forvaltningsafgørelser hos Jobcenter Vejen, hvad angår sagerne i vores datasæt. Vi bemærker, at vores begrænsede datasæt ikke understøtter en entydig konklusion om GPT-4's kapacitet til at træffe forvaltningsafgørelser generelt. Yderligere undersøgelser er således nødvendige for at kunne konkludere mere entydigt ift. GPT-4's kapacitet. Dette er i tråd med vores eksplorative tilgang, hvor vi alene ønskede at undersøge, om det er relevant at undersøge store sprogmodellers kapacitet til at foretage forvaltningsafgørelser. Her kan vi svare entydigt bekræftende. Yderligere undersø-

gelser af store sprogmodellers kapacitet til at foretage forvaltningsafgørelser er relevante.

Vi vurderer, at vi ville kunne forbedre GPT-sagsbehandlers præstation, hvis vi ikke havde begrænset os selv til altid kun at bruge første svar og ikke havde afholdt os fra at give sagsspecifik instruktion. Disse begrænsninger ville en sagsbehandler i et jobcenter i praksis ikke være underlagt. Vores resultat kan på den baggrund betragtes som en konservativ test af GPT-4's kapacitet ift. at træffe forvaltningsafgørelser.

En række forbehold gør sig gældende i forhold til vores resultat. For det første er vores stikprøvestørrelse begrænset. Det er sandsynligt, at vi med en større stikprøve ville have fundet en signifikant forskel mellem bedømmelsen af GPT-sagsbehandlers afgørelser og jobcenterets afgørelser. Det er dog, jf. tabel 3, ikke givet, i hvilken retningen forskellen i bedømmelserne ville være signifikant. Et resultat, der viste, at GPT-sagsbehandlers afgørelser bedømmes signifikant mere positivt end jobcenterets afgørelser, ville således være mindst lige så interessant som vores aktuelle resultat. Et andet forbehold vi må tage er, at GPT-sagsbehandlers afgørelser baserer sig på allerede fremstillet sagsmateriale, hvor sagsbehandlerne hos Jobcenter Vejen i kraft af deres foreløbige noter mv. kan tænkes at give GPT-sagsbehandler svar som GPT-sagsbehandler ellers ikke selv ville være fremkommet med. Et tredje forbehold gælder kvaliteten af Jobcenter Vejens afgørelser. Vores konklusion om, at bedømmelsen af GPT-sagsbehandlers afgørelser ikke er signifikant forskellig fra bedømmelsen af jobcenterets afgørelser, er således kun interessant i det omfang, at vi antager, at jobcenterets afgørelser er kvalificerede. Vi har ikke anledning til at antage anderledes, men vi har ikke foretaget en undersøgelse af denne antagelse.

Et forbehold vi ikke behøver at tage er, hvorvidt disse afgørelser er en del af GPT-4's træningsdata. Vi har gennemført vores eksperiment på materiale, der ikke er offentligt tilgængeligt, og vi kan derfor være ganske sikre på, at de korrekte afgørelser ikke er del af den data, som GPT-4 er trænet på. Dvs. vi kan altså udelukke, at GPT-sagsbehandlers svarkvalitet er kunstigt høj på den baggrund. Vi har i den forbindelse indstillet vores konto hos OpenAI således, at de data, vi indlæser, ikke bliver brugt til træning af fremtidige modeller.

Hvis store sprogmodeller på nuværende tidspunkt er i stand til at producere tilstrækkelige forvaltningsafgørelser, er det interessant af flere årsager. I det følgende vil vi kort diskutere nogle af implikationerne.

Implikationer

I National strategi for kunstig intelligens pointeres, at "[k]unstig intelligens skal hjælpe os med at analysere, forstå og træffe bedre beslutninger (Finansministeriet og Erhvervsministeriet, 2019). Men teknologien kan ikke – og skal ikke – erstatte mennesker eller træffe beslutninger for os." Dette udsagn er –

tyder vores undersøgelse på – allerede blevet overhalet af virkeligheden. Teknologien kan træffe beslutninger for os. Spørgsmålet er nu alene, om den skal.

Det kan forventes, at udviklingen på området vil fortsætte, og altså at de offentligt tilgængelige store sprogmodeller vil forbedres yderligere (se f.eks. Aschenbrenner, 2024). Når en offentlig tilgængelig sprogmodeller allerede i dag (medio 2024) er i stand til at producere afgørelser på niveau med et jobcenters faktiske afgørelser, er det således berettiget at have en forventning om, at fremtidige sprogmodeller vil kunne producere afgørelser på et niveau, der overstiger jobcentrenes faktiske afgørelser. Vi bemærker i den forbindelse, at Claude Opus 3's – Claude Opus 3 er en nyere model end GPT-4 – afgørelse (vedrørende M2 – S2) vurderes mere positivt end GPT-sagsbehandlers afgørelse af to ud af tre dommere.

Herved vil spørgsmålet om, hvorvidt store sprogmodeller skal benyttes til afgørelsesvirksomhed, blive tiltagende mere presserende, og det er på den lange bane svært for os at forestille sig, at en løsning, der er både billigere og bedre, ikke vil tiltrække sig lovgivers opmærksomhed. Dvs. spørgsmålene om rimelighed og lovlighed – jf. vores indledning – er praktisk relevante. Disse spørgsmål tager vi derfor hul på nedenfor. Denne diskussion af rimelighed og lovlighed er ikke et forsøg på at give endegyldige svar, men derimod et forsøg på at skitsere relevante elementer af en diskussion, der vil blive tiltagende aktuel.

Demokratiske implikationer

I Danmark og Skandinavien er der tradition for, at digitalisering er borgercentret (Aanestad o.a., 2024: 151), men denne tradition vil blive udfordret af store sprogmodellers fremmarch, idet disse modeller i deres kerne er uigennemskuelige (Wood o.a., 2024: 6). Denne iboende uigennemskuelige vil uundgåeligt bestyrke den allerede eksisterende magt- og informationsasymmetrien mellem offentlig forvaltning og borger, idet forvaltningen ikke vil være i stand til fyldestgørende at kunne forklare, hvorfor en given beslutning er blevet truffet. Forvaltningen kan forsøge at løse dette problem ved at "regne baglæns" og tilknytte allerede trufne beslutninger en forklaring evt. med store sprogmodellers mellemkomst, men sådanne post hoc-forklaringer vil ikke nødvendigvis være i overensstemmelse med den faktiske forklaring, der vil have sit sæde i interaktionen mellem input og en given sprogmodels parametre, hvilke skal tælles i billioner (engelsk trillions). Herved kan store sprogmodeller være med til at underminere tilliden mellem borgerne og forvaltningen, og også i sidste ende den demokratiske proces (Sigfrid o.a., 2023: 5). Helt konkret ved vi således heller ikke i vores eksperiment, hvorfor vores GPT-sagsbehandler svarer, som den svarer, og vi kan videre observere, at dens svar ikke nødvendigvis er konsistente over flere sessioner. Som bruger oplever vi altså modellens uigennemskuelighed, og vi vil ikke være i stand til fyldestgørende at forklare modellens afgørelser til en berørt boger eller på kvalificeret vis tilrettet modellens tilgang på baggrund af en politisk beslutning.

Etiske implikationer

Etisk er situationen også uklar. De åbenlyse faldgruber er risikoen for bias (Ferrara, 2023) og uigennemsigtighed (Wood o.a., 2024: 6) i grundlaget hvorpå sprogmodellen træffer afgørelser. Det er således åbenlyst, at brugen af ChatGPT til afgørelsesvirksomhed ikke lever op til Dataetisk Råds definition af dataetik. Dataetisk Råd har således formuleret ti principper og værdier, som er velfærd, værdighed, privatliv, selvbestemmelse, lighed, frihed, retssikkerhed, gennemsigtighed, sikkerhed og ansvarlighed (Dataetisk Råd, u.å.). Disse lever ChatGPT ikke op til. Det er ingen gennemsigtighed og ingen retssikkerheden, idet ChatGPT er en black box for danske myndigheder, der ikke har adgang til OpenAI's lukkede systemer. Det er heller ikke klart, at lighed sikres af systemet, idet bias overfor særlige grupper ikke kan udelukkes (Ferrara, 2023). Derudover kan det diskuteres, om et lukket system kan sikre værdighed og selvbestemmelse. Olsen (2023) argumenterer tilsvarende for, at black box-systemer har dataetiske udfordringer (Olsen, 2023: 118).

Det er dog også omvendt klart, at det ikke er uden etiske implikationer at fravælge en løsning, der potentielt kan levere afgørelser, fagpersoner vurderer som værende tilstrækkelige, og som også er billigere end alternative måder at producere afgørelser på. Olsen argumenterer således også for, at andre interesser skal vægtes end kun dataetiske overvejelser (Olsen, 2023: 118). Og dette er ikke nyt for store sprogmodeller. Kristensen beskriver således, hvordan "[g]*evinsterne ved digitalisering kommer [...] med en pris i form af en risiko for at kompromittere offentlige værdier som fx åbenhed og værdighed* " (Kristensen, 2023: 245). Sigfrid argumenterer for, at den pris er værd at diskutere, og at man skal overveje at turde sætte teknologierne fri og diskutere konsekvenserne i et bredere perspektiv end blot ud fra den borgercentrede vinkel (Sigfrid o.a., 2023: 6). Herved peger den etiske diskussion om brugen af store sprogmodellen tilbage på en klassisk diskussion om utilitaristisk etik (skal vi effektivere mest muligt for at skabe mest muligt nytte uden at tage hensyn til individet?) overfor deontologisk etisk (skal vi beskytte individets rettigheder uden hensyntagen til samfundsøkonomiske konsekvenser?). Dette er en diskussion, der ligger udenfor rammerne af denne artikel.

Juridiske implikationer

I en dansk forvaltningssammenhæng må det som allerede gennemgået anses som værende i modstrid med de almindelig forvaltningsretlige principper at benytte store sprogmodeller til afgørelsesvirksomhed, medmindre lovgiver har givet særskilt lovhjemmel hertil (jf. Loiborg, 2020). Det er også givet, at en bred vifte af de mulige anvendelsesområder vil være i modstrid med eller bliver indskrænket betydeligt af EU-lovgivning herunder både databeskyttelsesforordningen (Rasmussen og Klit, 2020) og AI-forordningen, hvor særligt artikel 14 om menneskeligt tilsyn og artikel 86 om retten til at få en forklaring om individuel beslutningstagning (Forordningen om kunstig intelligens, 2024) væsentlig begrænser brugen af kunstig intelligens særligt i form af store sprogmodeller i kraft af sprogmodellers iboende uigennemsigtighed (Wood

o.a., 2024: 6) . Samlet kan vi således med ganske høj sikkerhed konkludere, at myndigheder ikke uden lovgivers mellemvirke kan igangsætte en omstilling af deres forvaltningsvirksomhed til at være båret af kunstig intelligens i form af store sprogmodeller. Hvorvidt store sprogmodeller kan benyttes beslutningsunderstøttende uden lovgivers mellemvirke, er også uklart, idet en forvaltning ved anvendelse af en kommerciel model (som f.eks. GPT-4 eller Claude Opus 3) ikke vil have styring og kontrol med modellen (se evt. Loiborg, 2020: 71).

> I en dansk forvaltningssammenhæng må det som allerede gennemgået anses som værende i modstrid med de almindelig forvaltningsretlige principper at benytte store sprogmodeller til afgørelsesvirksomhed, medmindre lovgiver har givet særskilt lovhjemmel hertil

GPT-4 kan producere tilstrækkelige forvaltningsafgørelser

I denne artikel har vi undersøgt, om ChatGPT 4 (GPT-4) kan producere afgørelser, der er tilstrækkelige som forvaltningsafgørelser hos Jobcenter Vejen. Vores undersøgelse viser, at GPT-4 med et enkelt standardinstrukssæt kan producere forvaltningsafgørelser på baggrund af indlæst sagsmateriale, der ikke af vores dommerpanel bedømmes signifikant forskelligt fra de faktiske afgørelser, som Jobcenter Vejen har udformet. To dommere har dog hver ét spørgsmål, hvor vi med over 95 pct. sikkerhed kan konkludere, at der er forskel i deres bedømmelse af henholdsvis afgørelser produceret af ChatGPT 4 og Jobcenter Vejen.

Vi konkluderer på den baggrund bekræftende i forhold til vores undersøgelsesspørgsmål – dvs. vi konkluderer, at GPT-4 kan producere afgørelser, der er tilstrækkelige som forvaltningsafgørelser hos Jobcenter Vejen, hvad angår sagerne i vores datasæt. Vi bemærker, at vores begrænsede datasæt ikke understøtter en entydig konklusion om GPT-4's kapacitet til at træffe forvaltningsafgørelser. Yderligere undersøgelser er således nødvendige for at kunne konkludere mere entydigt ift. GPT-4's kapacitet. Dette er i tråd med vores eksplorative tilgang, hvor vi alene ønskede at undersøge, om det er relevant at undersøge store sprogmodellers kapacitet til at foretage forvaltningsafgørelser. Her kan vi konkludere entydigt bekræftende. Yderligere undersøgelser af store sprogmodellers kapacitet til at foretage forvaltningsafgørelser er relevante.

På baggrund af vores konklusion diskuterer vi kort forskellige implikationer af store sprogmodellers potentielle kapacitet til at foretage forvaltningsafgørelser. Vi finder, at det vil have demokratiske og etiske implikationer at implementere store sprogmodeller til at foretage forvaltningsafgørelser. Det er således et grundlæggende demokratisk problem, at store sprogmodeller i kraft

af deres arkitektur er uigennemskuelige, idet dette vanskeliggør demokratisk kontrol. Denne uigennemskuelighed udgør også et etisk problem, særligt når den kombineres med risikoen for indbygget bias. Omvendt finder vi også, at det ikke er uden etiske implikationer at undlade at bruge teknologi, der vil kunne effektivisere arbejdsgange. Endelig finder vi, at store sprogmodeller ikke umiddelbart kan implementeres i myndigheders afgørelsesvirksomhed uden lovgivers mellemkomst.

Noter

1 Hvorvidt denne brug følger almindelige forvaltningsretlige principper, og hvorvidt der er taget højde for potentielle etiske dilemmaer i den allerede foretagne implementering, falder udenfor denne artikels ramme.

Referencer

Aanestad, Margunn, Brit Ross Winthereik og Åsa Mäkitalo (2024), *Digital inclusion*, Djøf Forlag.

Anthropic (2024), "Introducing the next generation of Claude", www.anthropic.com/news/claude-3-family.

Aschenbrenner (2024), "I. From GPT-4 to AGI: Counting the OOMs – SITUATIONAL AWARENESS", https://situational-awareness.ai/from-gpt-4-to-agi/.

Brown, T.B. o.a. (2020), "Language Models are Few-Shot Learners", *Neural Information Processing Systems*, 33, 1877–1901, https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Bubeck, S. o.a. (2023), "Sparks of Artificial General Intelligence: Early experiments with GPT-4", *arXiv* (Cornell University), [Preprint], https://doi.org/10.48550/arxiv.2303.12712.

Dataetisk Råd (u.å.), *Om dataetik*, https://dataetiskraad.dk/om-dataetik.

Ekambaram, V. o.a. (2024), "Tiny Time Mixers (TTMs): fast pretrained models for enhanced Zero/Few-Shot forecasting of multivariate time series", *arXiv* (Cornell University) [Preprint], https://doi.org/10.48550/arxiv.2401.03955.

Faria-E-Castro, M. og F. Leibovici (2023), *Artificial intelligence and inflation forecasts*, https://doi.org/10.20955/wp.2023.015.

Ferrara, Emilio (2023), *Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models*, [Preprint], https://arxiv.org/pdf/2304.03738.

Finansministeriet og Erhvervsministeriet (2019), *National strategi for kunstig intelligens*, digst.dk.

Forordningen om kunstig intelligens: 2024/1689 (2024), https://eur-lex.europa.eu/legal-content/DA/TXT/HTML/?uri=OJ:L_202401689.

Jiang, K. og X. Lu (2020), "Natural Language Processing and its Applications in Machine Translation: A Diachronic review", *2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI)* [Preprint], https://doi.org/10.1109/iicspi51290.2020.9332458.

Kang, W.-C. o.a. (2023), "Do LLMs understand user preferences? Evaluating LLMs on user rating prediction", *arXiv* (Cornell University) [Preprint], https://doi.org/10.48550/arxiv.2305.06474.

KL og Digitaliseringsstyrelsen (2024), *Inspirationskatalog*, KL.dk. https://videncenter.kl.dk/media/fwehh4l0/inspirationskatalog-syv-eksempler-med-generativ-ai-i-den-offentlige-sektor.pdf.

Kristensen, K. (2023), "Etiske perspektiver på offentlig topledelse af digitalisering", *Politica*, 55(3): 242–51, https://doi.org/10.7146/politica.v55i3.140274.

Liang, Y. o.a. (2023), "Exploring Large Language Models for Human Mobility Prediction under Public Events", *arXiv* (Cornell University) [Preprint], https://doi.org/10.48550/arxiv.2311.17351.

LMSYS Chatbot Arena Leaderboard (2024), https://lmarena.ai/?leaderboard.

Loiborg, Emilie (2020), "Om den forvaltningsretlige ramme for overladelse af forvaltningsopgaver til systemer baseret på kunstig intelligens", *Ugeskrift for Retsvæsen*, 8, 64–72.

Lopez-Lira, A. og Y. Tang (2023), "Can ChatGPT forecast stock price movements? return predictability and large

language models", *Social Science Research Network* [Preprint], https://doi.org/10.2139/ssrn.4412788.

Mehdi, Y. (2024), "Bringing the full power of Copilot to more people and businesses – The Official Microsoft Blog", https://blogs.microsoft.com/blog/2024/01/15/bringing-the-full-power-of-copilot-to-more-people-and-businesses/.

NCBE (u.å.), "About the UBE | NCBE", www.ncbex.org/exams/ube/about-ube.

Olsen, B.K. (2023), *Håndbog i dataetik: Mennesker, miljø og samfund*, Hans Reitzels Forlag.

OpenAI (2023), "GPT-4 Technical Report", *arXiv* (Cornell University) [Preprint], https://doi.org/10.48550/arxiv.2303.08774.

Rasmussen, Nina og Mette Klit (2020), "Brug af Big Data og kunstig intelligens i den offentlige forvaltning", i Jesper Hundebøl, Anja Pors og Lars Sørensen, red., *Digitalisering i offentlig forvaltning*, 1. udg. Samfundslitteratur, pp. 223–42.

Schönegger, P. og P. S. Park (2023), "Large Language Model Prediction Capabilities: Evidence from a Real-World Forecasting Tournament", *arXiv* (Cornell University) [Preprint], https://doi.org/10.48550/arxiv.2310.13014.

Sigfrids, A. o.a. (2023), "Human-centricity in AI governance: A systemic approach", *Frontiers in Artificial Intelligence*, 6. https://doi.org/10.3389/frai.2023.976887.

Smith, Moloney og Asher-Schapiro (2023), "Are AI chatbots in courts putting justice at risk?", *Reuters*, 4. Maj, www.reuters.com/article/global-tech-justice-idUSL8N-36V3QL/.

Staff (2023), "ChatGPT-4 used in a Pakistani judgment as an experiment – courting the law", https://courtingthelaw.com/2023/04/07/laws-judgments-2/chatgpt-4-used-in-a-pakistani-judgment-as-an-experiment/.

Taylor, L. (2023), "Colombian judge says he used ChatGPT in ruling", *The Guardian*, 28. februar. www.theguardian.com/technology/2023/feb/03/colombia-judge-chatgpt-ruling.

Wood, Nathan, Kristian González Barman og Pawel Pawlowski (2024), "Beyond transparency and explainability: on the need for adequate and contextualized user guidelines for LLM use", *Ethics and Information Technology*, 26(47): 1–12, https://doi.org/10.1007/s10676-024-09778-2.