

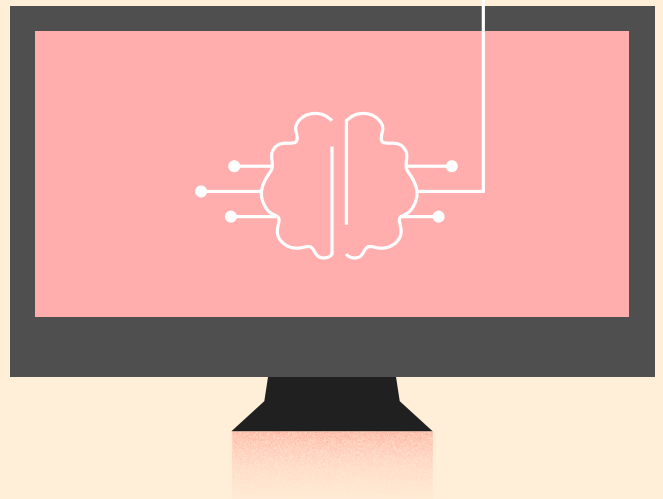
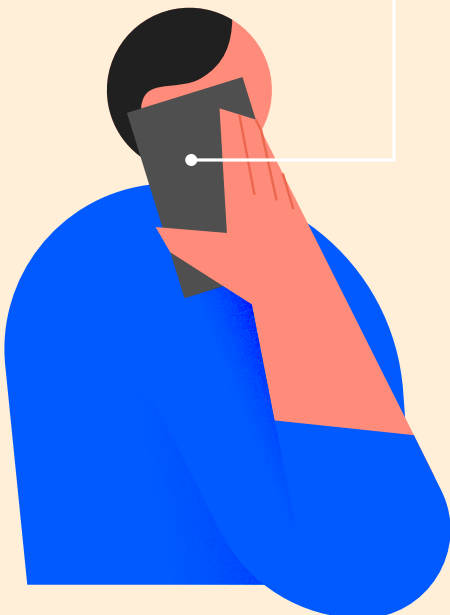


Dataetisk
Råd



Generativ AI og offentlige myndigheder

Pålidelighed og dataetik



Dataetisk Råd

København 2025.

Dataetisk Råd er en offentlig myndighed, som rådgiver om etiske udfordringer ved nye digitale teknologier. Læs mere på dataetiskraad.dk

Projektet bag denne rapport har i rådet været forankret i en referencegruppe bestående af:

- Anette Høyrup,
- Birgitte Arent Eiriksson,
- Christiane Vejøl,
- Eva Flyvholm,
- Peter Damm og
- Rikke Frank Jørgensen.

Projektet er udarbejdet med sekretariatsbistand fra projektleder Frej Klem Thomsen.

Layout og illustrationer ved Grobowski.

Rapporten er korrekturlæst med generativ AI.

Forslag til korrektur fra den kunstige intelligens indeholdt 19 fejl.

Indholdsfortegnelse

Indledning	6
1. Hvad er generativ AI?	10
1.1. Hvad er forskellen på generativ AI og andre former for kunstig intelligens?	15
1.2. Hvordan udvikler man generativ AI?	18
1.3. Hvordan virker generativ AI?	20
2. Generelle dataetiske udfordringer ved generativ AI	28
2.1. Udfordringer i anvendelse af generativ AI	29
2.2. Udfordringer med potentiale for misbrug af generativ AI	30
2.3. Udfordringer ved bredere effekter af generativ AI	32
2.4. Hvilke dataetiske udfordringer er særligt relevante for offentlige myndigheders anvendelse af generativ AI?	34
3. Generativ AI ved danske offentlige myndigheder	36
3.1. Kilder til Dataetisk Råds overblik	36
3.2. Hvad er det nye ved generativ AI?	38
3.3. Hvordan bruges generativ AI af offentlige myndigheder?	40
4. Eksisterende dataetiske retningslinjer for offentlige myndigheders brug af generativ AI	48
5. Fejl og kvalitetssikring	52
5.1. Hvor stor er risikoen for fejl i generativ AI?	55
5.2. Hvor stor en dataetisk udfordring udgør risikoen for fejl?	59
5.3. Hvad er den dataetisk acceptable risiko for fejl?	61
5.4. Kvalitetssikring af generativ AI	66
5.5. Hvor gode er eksisterende metoder til at kvalitetssikre generativ AI?	70
5.6. Menneskelig kvalitetssikring af generativ AI	71
5.7. Hvad er det dataetisk nødvendige niveau af kvalitetssikring?	73

6.	Risiko for fejl i offentlige myndigheders anvendelse af generativ AI: Dataetisk Råds position og anbefalinger	78
6.1.	Dataetiske overvejelser før anvendelse af generativ AI	79
6.2.	Dataetiske tiltag ved anvendelse af generativ AI	82
6.3.	Politiske initiativer for offentlige myndigheders anvendelse af GenAI	84



Indledning

Det er blevet en velkendt historie: I november 2022 lancerede OpenAI den generative AI ChatGPT-3.5. ChatGPT-3.5 satte en ny standard for, hvordan kunstig intelligens kunne interagere med og hjælpe brugere, ved at skabe tekst. Brugere kunne skrive med chatbotten og få svar, som i slående grad virkede som om, de var skrevet af et menneske. Chatbotten kunne svare på alskens spørgsmål, skrive sange, digte og tekster, generere computerkode, og lave korrekturlæsning. På få måneder fik den over 100 millioner registrerede brugere, og satte dermed rekord som den hurtigst voksende digitale tjeneste nogensinde. Siden da er generativ AI blevet udbredt med eksplosiv fart. I dag findes chatbots med generativ AI både frit tilgængelige på nettet og i mange kommercielle digitale tjenester, og vi har vænnet os til, at vi som almindelige brugere på et øjeblik kan skabe tekst eller naturligt udseende billeder med kunstig intelligens.

Vi bryster os i Danmark af, at vores offentlige sektor er verdens mest digitaliserede. Derfor er det ikke en overraskelse, at myndigheder har været hurtige til at gribe den nye teknologi, og udforske de muligheder den giver. Men med nye teknologier kommer uundgåeligt også nye etiske udfordringer.

Dataetisk Råd besluttede i 2024 at igangsætte et projekt, som skulle undersøge en særlig dataetisk udfordring ved offentlige myndigheders brug af generativ AI: risikoen for at generativ AI laver fejl, når den skaber indhold. Projektets resultater sammenfattes i en redegørelse, der udgør de første kapitler af denne rapport, og i Dataetisk Råds otte anbefalinger, rettet mod offentlige myndigheder og politiske beslutningstagere.

Rapporten begynder med at præsentere generativ AI som teknologi (kapitel 1). Kunstig intelligens er en kompleks teknologi, og kapitlet giver en generel indføring i hvordan generativ AI udvikles og fungerer, med særligt fokus på visse tekniske pointer og begreber, som er relevante for at forstå, hvorfor teknologien kan lave fejl, og hvordan offentlige myndigheder kan forholde sig til risikoen for fejl.

Generativ AI har allerede været genstand for intens debat og kritisk opmærksomhed i både den danske og internationale offentlige debat. Dataetisk Råd har i dette projekt fokuseret på den dataetiske udfordring med risiko for fejl, men det er vigtigt at forstå denne udfordring i kontekst af de andre dataetiske udfordringer, som generativ AI kan rejse. Rapporten giver derfor et overblik over de vigtigste dataetiske udfordringer (kapitel 2), og diskuterer hvordan forskellige typer udfordringer kan være særligt relevante for forskellige aktører.

Rådet har som nævnt også valgt at fokusere på netop offentlige myndigheders anvendelse af generativ AI. Offentlige myndigheder har et særligt ansvar for at informere korrekt og træffe korrekte beslutninger, som skyldes den effekt fejlagtig information og fejlagtige beslutninger kan have for borgere, samfund og miljø. For at forstå udfordringen i denne sammenhæng har rådet undersøgt, hvordan generativ AI anvendes og kan anvendes af offentlige myndigheder i Danmark. Rapporten præsenterer seks forskellige brugsscenerier (kapitel 3), og diskuterer hvordan risikoen for fejl varierer i de forskellige scenarier.

En række aktører har allerede udarbejdet retningslinjer for offentlige myndigheders arbejde med generativ AI, ligesom både danske og internationale aktører har diskuteret de dataetiske udfordringer ved anvendelse af generativ AI, og foreslået tiltag til at tackle disse udfordringer. Rapporten sammenfatter de mest almindelige anbefalinger fra en række danske og internationale retningslinjer (kapitel 4), og diskuterer styrker og svagheder ved disse retningslinjer.

Analysen af eksisterende retningslinjer er udgangspunktet for diskussion af den centrale dataetiske udfordring med risiko for fejl, når offentlige myndigheder anvender generativ AI (kapitel 5). Dataetisk Råds analyse peger på, at risikoen for fejl ikke kan elimineres, men at forskellige brugsscenerier kan have forskellig risikoprofil. I nogle brugsscenerier laver generativ AI flere fejl, og disse fejl kan have afgørende betydning for borgere og medarbejdere; i andre brugsscenerier laver generativ AI færre fejl, og disse fejl har lille eller ingen betydning for borgere og medarbejdere. Det viser, at det er vigtigt, at vurdere generativ Als risikoprofil i det enkelte brugsscenarie. Imidlertid er det teknisk vanskeligt, at kvalitetssikre generativ AI ved at måle, hvor mange fejl et system laver. Ofte er den bedste metode at udføre menneskelige test, hvor medarbejdere vurderer mange eksempler på indhold skabt med generativ AI. Disse forhold rejser tilsammen to dataetiske kernespørgsmål, om hvornår risikoen for fejl bliver så høj, at offentlige myndigheder bør undlade at anvende et generativ AI-system, og om hvor sikker en offentlig myndighed skal være på, hvilken risikoprofil generativ AI har.

På baggrund af rapportens redegørelse for teknologien, generelle dataetiske udfordringer, anvendelsen ved offentlige myndigheder, eksisterende dataetiske retningslinjer, og den særlige dataetiske udfordring med risiko for fejl, tager Dataetisk Råd stilling til, hvordan offentlige myndigheder bør forholde sig til udfordringen (kapitel 6). Rådets

position sammenfattes i otte anbefalinger, som både retter sig mod den enkelte myndighed, der ønsker at anvende generativ AI, og mod behovet for nye politiske tiltag.

Dataetisk Råds anbefalinger

Det er afgørende, at vi som samfund forholder os til de dataetiske udfordringer, som udvikling og anvendelse af GenAI rejser. Offentlige myndigheder, som anvender GenAI, har et særligt ansvar for at tackle disse dataetiske udfordringer.

- 1. Offentlige myndigheder** bør begynde eventuel udforskning og anvendelse af GenAI med brugsscenarier, hvor anvendelsen har lav risikoprofil, og inddrage interessenter i udvikling og evaluering af systemer.
- 2. Offentlige myndigheder** bør skelne mellem GenAI-systemer med høj og lav risikoprofil, og kun bruge GenAI, når myndigheden har indført tilstrækkelige mitigerende tiltag til at begrænse risikoen for fejl.
- 3. Offentlige myndigheder** bør systematisk vurdere GenAI-systemers samlede risikoprofil for fejl i hvert enkelt brugsscenarie.
- 4. Offentlige myndigheder** bør gennemføre mere omfattende kvalitets sikring af et GenAI-system, jo højere systemets risikoprofil er.
- 5. Offentlige myndigheder** bør oplyse, hvordan de anvender menneskelig kvalitetskontrol af GenAI, og tildele relevante medarbejdere de nødvendige ressourcer.
- 6. Offentlige myndigheder** bør indføre nødvendige forholdsregler til at identificere, dokumentere, oplyse og korrigere fejl, inden de tager et GenAI-system i brug.
- 7. Regeringen** bør udvikle fælles retningslinjer for offentlige myndigheds anvendelse af GenAI, som slår fast hvordan myndigheder skal håndtere risikoen for fejl.
- 8. Regeringen** bør oprette et nationalt, offentligt register over GenAI-systemer, som anvendes af offentlige myndigheder.

Dataetisk Råd har under arbejdet haft stor gavn af dialog med eksperter, myndigheder og interessenter. Rådet skylder en særlig tak til ATP, Digitaliseringsstyrelsen, fagdirektør Finn Myrstad og det norske Forbrukerrådet, Gladsaxe kommune, KLS Videncenter for digitalisering og teknologi, politisk direktør Mikael Ekman og CTO Rasmus Knappe fra Microsoft Danmark, Miljøstyrelsen, Professor Ole Winther fra DTU og Københavns Universitet, CSO Toine Bogers fra Pionercentret for Kunstig Intelligens og deltagerne ved Dataetisk Råds workshop på D3A-2024-konferencen, Lektor Tommy Sonne Alstrøm fra DTU, Vejle kommune, Lektor Zuzanna Godzimirska og AI & Human Rights-projektet ved Københavns Universitet, og Århus Kommune.

Dataetisk Råd har valgt at tage fat på arbejdet med generativ AI ved at kigge på den dataetiske udfordring med risikoen for fejl, når generativ AI anvendes af offentlige myndigheder, fordi det er en hjørnesten i vores samfund, at borgere kan have tillid til, at myndigheder informerer og sagsbehandler korrekt. Rådet er stolt af at kunne præsentere en af de første dataetiske analyser af denne udfordring i både dansk og international sammenhæng. Men generativ AI er en banebrydende teknologi, som kommer til at have stor indflydelse på vores samfund i de kommende år, og som rejser mange komplekse dataetiske udfordringer. Det er vigtigt, at vi i Danmark tager hul på at diskutere og forholde os til disse mange udfordringer, så vi i fællesskab sikrer, at vi udvikler og anvender generativ AI på en måde, hvor vi på én gang høster gevinsterne, og undgår de etiske faldgruber. Dataetisk Råd ser frem til også i de kommende år, at være en del af den fælles samtale om, hvordan vi skal løse disse udfordringer.



Johan Busse

Formand for Dataetisk Råd

8. januar 2025

1. Hvad er generativ AI?

Generativ kunstig intelligens (AI) refererer til en gren af kunstig intelligens, der omhandler design og implementering af algoritmer, som kan generere nyt indhold. Dette indhold kan variere fra tekst, billeder, musik, til video, der ofte er af en sådan kvalitet, at det kunne antages at være skabt af mennesker. Generativ AI adskiller sig fra traditionel AI, der typisk fokuserer på at analysere og fortolke data, ved i stedet at skabe nyt og originalt indhold baseret på de data, den er trænet på.

I hjertet af generativ AI ligger maskinlæringsmodeller og især en underkategori kendt som dyb læring. Disse modeller er baseret på kunstige neurale netværk, der efterligner den menneskelige hjernes struktur og funktion. Ved at træne disse modeller på omfattende datasæt kan de lære at genkende mønstre, strukturer og forhold i dataene, hvilket giver dem evnen til at generere nyt indhold, der afspejler det, de er trænet på.

En fremtrædende anvendelse af generativ AI er inden for kreative felter, såsom kunst, musik og design, hvor det bruges til at skabe nye værker og design, der kan udfordre eller udvide de eksisterende normer inden for disse områder. For eksempel kan kunstnere anvende generativ AI til at generere nye kunstværker, mens komponister kan bruge teknologien til at skabe nye musikstykker.

Ud over de kreative industrier finder generativ AI anvendelse i mange andre sektorer. I forretningsverdenen kan det bruges til at generere realistiske datasæt til træning af andre AI-modeller, udvikle forbedrede chatbots til kundeservice, eller skabe personaliserede uddannelsesmateriale. Det har også potentiale inden for spiludvikling, ved at skabe dynamiske spilomgivelser eller generere realistiske NPC'er (non-player characters).

Til trods for de mange fordele og anvendelser, indebærer brugen af generativ AI også visse udfordringer og etiske overvejelser. Disse omfatter spørgsmål omkring ophavsret, når AI skaber kunst eller musik, samt risici forbundet med misinformation eller deepfakes, hvor AI genererer realistiske, men →

falske billeder eller videoer. Derfor er det vigtigt, at der findes retningslinjer og regler for anvendelsen af generativ AI for at sikre, at teknologien anvendes ansvarligt og til samfundets bedste.

I resumé er generativ AI et voksende forskningsområde inden for kunstig intelligens, der udvider grænserne for, hvad computere kan skabe. Med sit potentiale til at revolutionere mange industrier, fra kunst til uddannelse og forretning, følger et ansvar for at navigere i de etiske og samfundsmæssige udfordringer, det bringer med sig.

Præsentation af generativ AI produceret af ChatGPT 4.0, 13. marts 2024. Prompt: "Skriv en 400-ord lang præsentation og definition af generativ AI rettet mod et bredt publikum. Anvend en nøgtern og ikke-vurderende tone – genren er et leksika-opslag ligesom Wikipedia."



Generativ AI (GenAI) er en særlig type kunstig intelligens, som er kendetegnet ved, at den kan skabe originalt indhold af høj kvalitet, for eksempel tekst, billeder eller lyd. Typisk kræver det blot et simpelt input fra brugeren, et såkaldt "prompt", at få en GenAI til at skabe indhold.¹

Prompt

Et simpelt brugerinput, i form af en besked eller et spørgsmål, som GenAI anvender til at generere indhold.

Fire almindelige former for GenAI er **tekst-generation, billedgeneration, lydgeneration** og **videogeneration**.

Tekstgeneration

Den mest udbredte og almindeligt kendte form for GenAI er tekstgeneration. En tekst-generator skaber ny tekst på baggrund af prompt fra brugeren.

Tekstgeneration er et fleksibelt digitalt værktøj og kan typisk skabe mange forskellige slags tekst, med den genre og det indhold, som brugeren ønsker. Den kan for eksempel generere sange, digte, nyhedsartikler, computerkode, læserbreve, opskrifter på mad, og analyser af et emne eller tekst. Og den kan revidere eller kommentere på tekstinput, for eksempel for at oversætte, finde fejl eller foreslå måder at forkorte og opsummere.

Tekstgeneration anvendes ofte i chatbots, hvor den kunstige intelligens simulerer den form for skriftlig interaktion, som mennesker har, når de skriver beskeder med hinanden. Sådanne chatbots kan eksempelvis vejlede kunder, patienter eller borgere på en hjemmeside for at hjælpe dem med at finde relevant information. Avancerede chatbots kan hjælpe brugeren med at udføre handlinger, for eksempel klage over en vare eller service som brugeren har købt, lave en aftale, eller downloade, udfylde og indsende en formular. Et omdiskuteret eksempel er digitale tjenester, som lader brugeren skabe en individuel AI-persona, der kan fungere som virtuel kæreste, ven eller mentor. Brugeren kan vælge karakteristika og egenskaber til sin persona, men den kunstige intelligens kan også lære af de samtaler, som den har med brugeren.

Kendte eksempler på GenAI til tekstgeneration inkluderer **Anthropics Claude, Googles Gemini, Microsofts CoPilot, og OpenAIs ChatGPT**.

¹ For overblik, se Jones, et al. 2023.

Billedgeneration

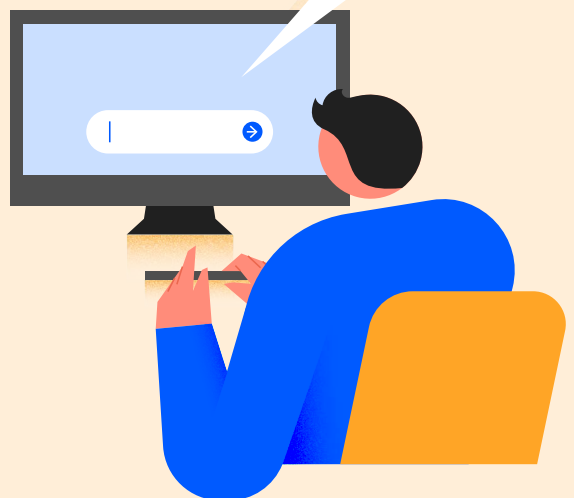
En anden populær form for GenAI er billedgeneration. En billedgenerator skaber billeder efter brugerens ønske baseret på simple tekstinput. Typisk behøver brugeren blot at skrive en prompt med en kort beskrivelse af det ønskede motiv, hvorpå GenAI skaber et billede, som passer til beskrivelsen. Men brugere kan også indtaste mere detaljerede beskrivelser af motivet, inklusive genre, vinkler, lyssætning og stemning.

Billedgeneration er teknisk en relativt vanskelig opgave, og GenAI producerer fortsat ofte billeder, som på forskellige måder ikke passer til brugerens ønsker, for eksempel urealistiske eller umulige motiver. Ikke desto mindre virker billedgeneration ofte så godt, at det producerer billeder, som for almindelige brugere er svære at skelne fra billeder skabt af menneskelige kunstnere eller fotografer.

Kendte eksempler på GenAI til billedgeneration inkluderer **Googles Imagen**, **Midjourney**, **OpenAIs DALL-E**, og **Stability AIs Stable Diffusion**.



Billeder genereret med DALL-E 3 med følgende prompt: "A modern-day middle-age academic working on her computer at her desk, writing a report on generative AI. The office environment is spacious and cozy, with bookshelves holding both books and art in the background. The style is a grainy black and white photo / a Japanese manga / an impressionist painting / a stick-figure pencil drawing."



Lydgeneration

En tredje udbredt form for GenAI er lydgeneration, hvor den kunstige intelligens skaber lyd i overensstemmelse med brugerens input. Den måske mest kendte anvendelse er generation af tale. For at skabe en lydfil med menneskelig tale behøver brugeren blot at fodre den kunstige intelligens med en tekst, og bestemme hvilken type stemme, som skal udtale teksten. Det kan være en GenAI-skabt stemme, som imiterer for eksempel et bestemt køn, en bestemt alder og en bestemt accent, men det kan også være en GenAI-imitation af en virkelig persons stemme.

Kreative brugere har for eksempel benyttet GenAI til at skabe stemmer der simulerer, at de amerikanske politikere Barack Obama, Joe Biden og Donald Trump spiller et multiplayer-computerspil sammen.² Andre har brugt GenAI til at få den afdøde sanger Johnny Cash til at synges cover-versioner af nye pophits.³

Kendte eksempler på GenAI til lydgeneration inkluderer **ElevenLabs** og **15.ai**.

Videogeneration

En naturlig forlængelse af billedgeneration er videogeneration, idet en video blot er en serie af billeder. Videogeneration har imidlertid vist sig, at være endnu mere teknisk krævende end billedgeneration, fordi det kræver, at GenAI kan skabe en naturlig sekvens mellem de enkelte billeder. De mest avancerede modeller i 2024 laver endnu så relativt mange fejl, at det typisk er let at se, at der er tale om indhold skabt af en GenAI.

Kendte eksempler på GenAI til videogeneration inkluderer **OpenAIs Sora**, **Metas Make-a-video** og **Runways Gen-3**.

² Se eksempelvis et spil Minecraft: <https://www.youtube.com/watch?v=-pGT5jLlx4>.

³ Se eksempelvis en cover-version af Taylor Swifts "Blank Space", eller et blues medley af blandt andet Aquas "Barbie Girl!": <https://www.youtube.com/shorts/k14Yx7qmGx4> og <https://www.youtube.com/watch?v=MAFdZBTø2lg>.

1.1. Hvad er forskellen på generativ AI og andre former for kunstig intelligens?

Generativ AI (GenAI) er en særlig form for kunstig intelligens, som har fået meget opmærksomhed de seneste år. Kunstig intelligens kan imidlertid fungere på mange måder, og også andre former for kunstig intelligens har de seneste år fyldt meget i den offentlige debat. Det skyldes at kunstig intelligens i stigende grad er en teknologi, som vi møder i vores hverdag. Almindelige eksempler inkluderer:

- **Spamfiltre**, som analyserer indkommende e-mail for at sortere uønskede mails fra.
- **Billedgenkendelse**, for automatisk at bestemme hvad et billede forestiller, og for eksempel samle alle billeder med et bestemt motiv.
- **Talegenkendelse**, som transskriberer tale til tekst, og for eksempel gør det muligt at diktere en besked til sin telefon.
- **Ansigtsgenkendelse**, som lader computeren identificere en person, og for eksempel gør det muligt, at låse en mobiltelefon op, ved at holde telefonens kamera foran sit ansigt.
- **Simulerede agenter**, som reagerer på den situation de er i, for eksempel karakterer i computerspil, som reagerer på spillerens handlinger, eller en robotstøvsuger, som ændrer retning, når den støder på en forhindring.
- **Rejseplanlægning**, som analyserer information om for eksempel transportmuligheder, geografi, trafikforhold og færdselspræferencer, for at planlægge en rute og vise vej.
- **Anbefalingsfiltre**, for eksempel på sociale medier og medieplatforme, som tilpasser det indhold den enkelte bruger bliver præsenteret for.
- **Automatiserede beslutningssystemer**, som vurderer en valgsituation og anbefaler en beslutning, for eksempel en diagnose i sundhedsvæsenet, eller en trivselsmåling på arbejdsmarkedet.

Selvom vi både anvender kunstig intelligens stadigt mere hyppigt, og ofte diskuterer emnet, kan det være svært at sige, præcis hvad kunstig intelligens er.⁴ Det skyldes dels at kunstig intelligens er en kompleks teknologi, og dels at vi ikke altid er sikre på eller enige om, hvorvidt en given teknologi skal betragtes som et eksempel på kunstig intelligens.

Mange formelle definitioner af kunstig intelligens er også relativt komplekse. EU's "Ekspertgruppe på højt niveau om kunstig intelligens" definerer eksempelvis kunstig intelligens ved at hævde at:

*"AI-systemer er softwaresystemer (og evt. også hardware-systemer), der er designet af mennesker, som ud fra et komplekst mål fungerer i den fysiske eller digitale dimension, hvor de opfatter deres miljø ved at indhente data, fortolker de indsamlede strukturerede eller ustrukturerede data, ræsonnerer ud fra forholdene eller behandler de oplysninger, der er udledt af disse data, og vælger den handling, der bedst kan opfylde det angivne mål. AI-systemer kan enten bruge symbolske regler eller lære en numerisk model, og de kan tilpasse deres adfærd ved at analysere, hvordan miljøet blev påvirket af deres tidligere handlinger."*⁵

Ekspertgruppe-definitionen har både styrker og svagheder. Definitionen er kompleks, og kan være en begrænset hjælp for personer med begrænset kendskab til teknologien, som ønsker bedre at forstå kunstig intelligens. Den synes også på nogle punkter at definere kunstig intelligens for snævert. Eksempelvis er det ikke klart, hvorfor kunstig intelligens skal være designet af mennesker – antageligt kunne software, som blev designet af en kunstig intelligens, selv være en kunstig intelligens. Det er heller ikke klart, at en kunstig intelligens nødvendigvis skal have "evnen til at tilpasse sin adfærd ved at analysere, hvordan miljøet blev påvirket af [...] tidligere handlinger", selvom nogle former for kunstig intelligens kan have denne evne.

4 Se McCarthy 2007; Wang 2019.

5 Den uafhængige ekspertgruppe på højt niveau om kunstig intelligens 2018.

I nogle situationer kan en simpel, intuitiv forståelse af "kunstig intelligens" være lettere at arbejde med. De fleste forbinder kunstig intelligens med computersystemer, som er i stand til at løse komplekse opgaver. Men vi skelner mellem forskellige typer opgaver. Vi kalder det ikke kunstig intelligens, når vi taler om for eksempel et styresystem til computere, eller den software som kontrollerer et digitalt kamera, selvom disse systemer løser meget komplekse opgaver. I denne rapport anvender Dataetisk Råd en bred og letforståelig definition af kunstig intelligens, som et computersystem, der løser en kompleks opgave, det ellers ville kræve menneskelig intelligens at løse.⁶

Kunstig intelligens

Et computersystem, som er i stand til at løse komplekse opgaver, det normalt ville kræve menneskelig intelligens at løse.

GenAI er ifølge denne definition de særlige former for kunstig intelligens, som løser opgaver med at skabe indhold, som det normalt ville kræve menneskelig intelligens at skabe.



6 For yderligere diskussion, se Bringsjord & Govindarajulu 2018.

1.2. Hvordan udvikler man generativ AI?

Kernen i alle former for kunstig intelligens er en matematisk model. Når man udvikler en kunstig intelligens, skal man bestemme strukturen og værdierne i denne matematiske model. Der findes grundlæggende to metoder til at udvikle kunstig intelligens: regelbaseret udvikling og maskinlæring.

Kunstig intelligens blev i det meste af det 20. århundrede udviklet ved såkaldt regelbaseret udvikling, hvor mennesker designer regler for systemet, ved manuelt at bestemme struktur og værdier i den matematiske model. For eksempel kan menneskelige skakeksperter samarbejde med programmører om at udvikle kunstig intelligens som kan spille skak, ved at formulere matematiske regler for, hvordan systemet skal evaluere brikernes positioner på brættet, og vælge et træk. Kunstig intelligens skabt på denne måde kan være meget stærk, simpelthen fordi computersystemer kan foretage komplekse matematiske beregninger langt hurtigere end mennesker. Det illustrerede IBM allerede i 1997, da skakcomputeren Deep Blue vandt over den daværende verdensmester i skak, Garry Kasparov. Men de beregninger systemet foretager er baseret på menneskers forståelse af hvordan systemet kan løse problemet, og principielt forståelige for mennesker.

De seneste årtier er det blevet stadig mere almindeligt, at udvikle kunstig intelligens ved at bruge såkaldt "maskinlæring". Ved maskinlæring bruger udvikleren ét computersystem – en læringsalgoritme – til at udvikle det andet system – den kunstige intelligens. Læringsalgoritmen fodres med data om den type opgave, som den kunstige intelligens skal løse, og analyserer disse data, for at lave en matematisk model for hvordan forskellige data hænger sammen med resultater. Det kaldes at læringsalgoritmen "træner" systemet. Kvaliteten af den kunstige intelligens afhænger blandt andet af, hvor mange relevante data det fodres med – jo mere relevant data, jo bedre bliver systemet typisk til at løse opgaven. Ved at lære af store mængder data får den kunstige intelligens mulighed for at forstå problemet på nye og originale måder, og finde løsninger, som mennesker ikke ville have fundet på at formulere som regler. Til gengæld er kunstig intelligens udviklet med maskinlæring ofte meget kompleks, og fungerer på måder, som kan være meget vanskelige eller endda umulige at gennemskue for mennesker.

Maskinlæring

En metode til at udvikle kunstig intelligens, ved at træne på store mængder data.

GenAI udvikles med maskinlæring. Modellen trænes på eksempler på tekst, billeder eller lyd, for at lære hvordan forskellige typer data hænger sammen. Modellens kvalitet afhænger af både mængden og kvaliteten af de træningsdata, som den kunstige intelligens lærer af. Mange former for kunstig intelligens kan trænes på nogle få hundrede eksempler, mens mere kompleks AI trænes på millioner eller milliarder af eksempler. Når udviklere træner GenAI bruger de typisk endnu flere data. Data specialister taler om, at de mest avancerede GenAI-modeller trænes på "en væsentlig bid af internettet". Det inkluderer ikke blot skønlitteratur og faglitteratur, men også avisartikler, hjemmesider, og indlæg på sociale medier og online-fora. OpenAIs GPT4.0 er angiveligt trænet på tekster med samlet cirka 13 trillioner tokens – det svarer til rundt regnet 100 millioner bøger.⁷ Træning af så avanceret kunstig intelligens kræver enorm regnekraft, og tager ofte dage eller uger, selv når den udføres på specialiserede supercomputere.

Kunstig intelligens

Et computersystem, som er i stand til at løse komplekse opgaver, det normalt ville kræve menneskelig intelligens at løse.

Maskinlæring

En metode til at udvikle kunstig intelligens, ved at træne på store mængder data.

Generativ AI

En maskinlært kunstig intelligens, som med simple brugerinput kan skabe originalt indhold af høj kvalitet.



7 Se Walker 2023; George 2023.

1.3. Hvordan virker generativ AI?

Kernen i mange former for GenAI er en såkaldt "stor sprogmodel" (eng. "large language model"), forkortet LLM.⁸ En LLM er en matematisk model for menneskeligt sprog, som statistisk viser, hvordan sproglige udtryk meningsfuldt hænger sammen.

LLM

En kompleks matematisk model for statistiske sammenhænge i menneskeligt sprog.

En LLM løser den tilsyneladende enkle opgave hele tiden at forudsige den næste del af en tekst. I det eksempel, som vi startede dette kapitel med, skriver brugeren til ChatGPT:

"Skriv en 400-ord lang præsentation og definition af generativ AI rettet mod et bredt publikum."

ChatGPT betragter denne kommando som begyndelsen af en tekst, og svaret som en fortsættelse. Modellen vurderer derfor, hvad den mest sandsynlige velfungerende fortsættelse af teksten er, og genererer denne fortsættelse bid for bid, indtil den har løst den overordnede opgave.

Token

Et sprogligt brudstykke, for eksempel en stavelse, et bogstav eller et tegn, som udgør de enkelte bidder af tekst, en LLM gradvist genererer.

Når en LLM genererer tekst, arbejder den med et ordforråd, der består sproglige brudstykker, som kaldes "tokens". En token kan være et ord eller et tegn, men også en almindelig kombination af tegn, for eksempel en stavelse. I eksemplet på tekstgeneration i begyndelsen af kapitlet er det første ord "Generativ". Her kunne "Gen", som

udgør begyndelsen af ordet "generativ" være en token, ligesom slutningen af ordet "-iv" kan være en token, fordi det er en relativt almindelig endelse.

⁸ Store sprogmodeller (LLM) er én variant af såkaldte grundlagsmodeller (eng. "foundation models") eller grænsemødelles (eng. "frontier models"). Se Jones, et al. 2023; Bommasani, et al. 2021. GenAI-systemer anvender ofte LLM i kombination med andre typer GenAI-modeller, for eksempel såkaldte diffusionsmodeller til billedgeneration. Selvom mange pointerer gælder for alle de forskellige typer GenAI, fokuserer vi her især på LLM.

En LLM gør det nemt at løse mange ellers vanskelige opgaver med kunstig intelligens. Det er eksempelvis relativt enkelt for en myndighed at udvikle en chatbot, som kan hjælpe med at besvare spørgsmål fra borgere, når chatbotten kan bruge en LLMs matematiske model for, hvordan menneskeligt sprog fungerer som grundlag.

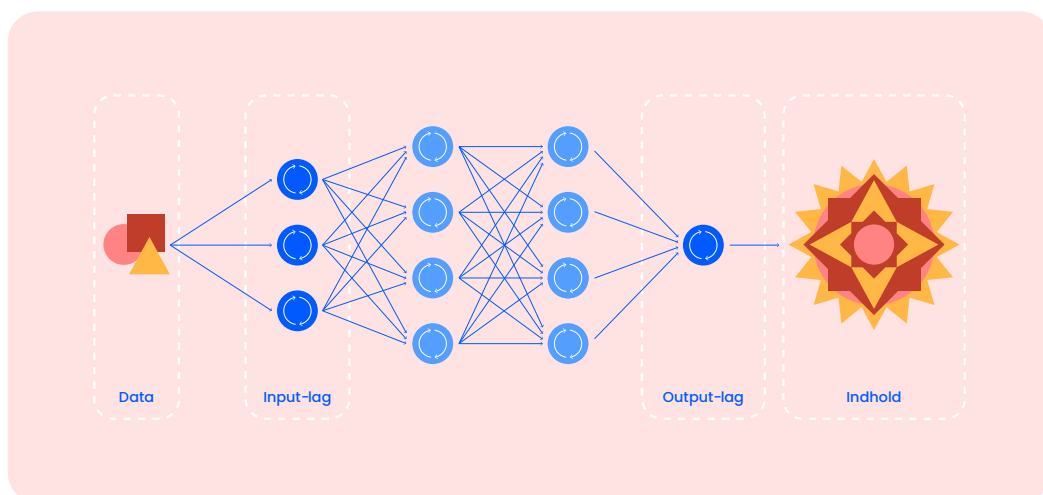
De matematiske modeller, som anvendes i GenAI, er en type kunstig intelligens, som kaldes for "dybe neurale netværk". Dybe neurale netværk er vundet frem de seneste årtier, fordi de ofte er i stand til at løse komplekse opgaver, som det ellers har været vanskeligt at løse med kunstig intelligens, for eksempel billedgenkendelse. Modellen kaldes for et neuralt netværk, fordi dens struktur imiterer den måde,

Neuralt netværk

En kunstig intelligens, hvor matematiske funktioner er organiseret i et netværk af "neuroner", der hver for sig behandler data, og sender dem videre i netværket.

neuronerne i den menneskelige hjerne er forbundet, og udveksler information. Et neuralt netværk består af et system af matematiske funktioner – digitale "neuroner" – som hver for sig behandler data, og udveksler data med hinanden. Et neuralt netværk er dybt, når der er mange lag af digitale neuroner, som er forbundet med hinanden.

Et neuralt netværk kan være mere eller mindre komplekst, afhængigt af hvor mange forbindelser mellem "neuroner" netværket består af, og hvordan de er organiseret. Et simpelt neuralt netværk kan, som i det illustrerede eksempel, have 32 forbindelser mellem neuroner organiseret i 4 lag.



GenAI hører til de dybeste, mest komplekse neurale netværk, mennesker har udviklet. Den frit tilgængelige model "Llama", som er udviklet af Meta, er et neuralt netværk, som i den mest avancerede version har 90 milliarder forbindelser mellem de digitale neuroner. OpenAIs GPT 4 model er ikke frit tilgængelig, men ansås til at være endnu større med cirka 1.700 milliarder forbindelser.⁹ Sådanne modeller er så komplekse, at de er uigennemskuelige for mennesker – selv udviklerne kan ikke i detaljer forstå, hvordan et prompt får modellen til at skabe indhold.

Kontekst og opmærksomhed

GenAIs evne til at skabe naturlig tekst ved at forudsige, hvordan teksten skal fortsætte, minder på flere måder om den "auto-complete"-funktion, som de fleste kender fra en moderne mobiltelefon. Auto-complete virker på den måde, at når brugeren har indtastet et ord, så foreslår funktionen en mindre gruppe ord, som kunne være de næste i beskeden.

Auto-complete gør det nemmere at skrive beskeder, men har den velkendte begrænsning, at det ofte foreslår de forkerte ord. Det skyldes blandt andet, at algoritmen kigger på en begrænset del af den sproglige kontekst. Den mest simple form for auto-complete kigger kun på det ene ord, som brugeren lige har skrevet, og foreslår ganske enkelt de ord, som hyppigst kommer efter netop dette ord på Dansk. Mere avancerede versioner kigger på lidt mere af den sproglige kontekst, som ordet optræder i, og bliver derved bedre til at gætte på, hvad det næste ord skal være. Men fordi auto-complete ikke vurderer den samlede tekst er den ude af stand til af sig selv, at skabe egentlig meningsfuld tekst.

Autocomplete

"At det ikke sikkert det lyder super god service til vores nabo at det ikke har noget med det formål har vi en række og jeg kan godt være en del på mine justerede at det ikke sikkert det rigtige valg..."

Et eksempel på en tekst, genereret ved konsekvent at acceptere det første forslag til det næste ord fra en auto-complete funktion.

⁹ Se Walker 2023.

Hvorfor vurderer auto-complete ikke hele konteksten? Det korte svar er, at det hurtigt bliver meget teknisk krævende, at få en kunstig intelligens til at vurdere, hvilket ord, som er den mest sandsynlige fortsættelse, hvis den skal forholde sig til hele den kontekst, som ordet optræder i. En oplagt løsning er at fokusere opmærksomheden på de dele af konteksten, som har mest betydning for forudsigelsen.

Den store succes, som GenAI har nydt i de seneste år, skyldes et teknisk gennembrud i 2017: transformer-arkitekturen.¹⁰ Transformeren er en software-struktur, der kan bygges ind i kunstig intelligens, som består af to dele: en indkoder og en dekoder. Indkoderen analyserer den sproglige kontekst, og oversætter information om tekstens indhold og struktur til matematiske værdier. Dekoderen bruger disse værdier til at hjælpe LLM-modellen, med at generere tekst. Ved at bruge transformer-arkitektur, kan en LLM fokusere opmærksomheden på de mest relevante dele af den sproglige kontekst. Tilsammen gør indkodning og dekodning det muligt for GenAI, at arbejde med en stor sproglig kontekst, og lægge vægt på netop de dele af konteksten, som er relevante for den token, den skal generere.¹¹

Opmærksomhed

"Skriv en 400-ord lang præsentation og definition af generativ AI rettet mod et bredt publikum. Anvend en nøgtern og ikke-vurderende tone – genren er et leksika-opslag ligesom Wikipedia."

Et eksempel på en prompt, med markering af nogle ord, som det kunne være vigtigt, at være opmærksom på, og andre ord eller tegn, som antageligt kan spille en mindre rolle for, hvordan svaret skal genereres.

Hallucinationer, temperatur og stokastiske papegøjer

GenAI er i stand til at skabe meget naturligt udseende tekst, og kan ofte besvare spørgsmål med detaljeret, præcis information. Det er derfor vigtigt at holde sig for øje, at en LLM ikke forstår hvad en tekst betyder – heller ikke den tekst den genererer. GenAI skaber tekst ved at gentage mønstre i de tekster den er trænet på. Skridt for

¹⁰ Se Vaswani 2017.

¹¹ Se Alammr 2018.

skridt genererer modellen en af de tokens, som den statistisk har lært, vil være vel-egnede fortsættelser.

Fordi en LLM ikke forstår hvad en tekst betyder, kan den heller ikke skelne mellem, om en tekst er faktisk korrekt eller faktisk forkert. En LLM kan generere velformuleret og meningsfuld tekst, som er fuld af fejlagtig information. Det kaldes at GenAI "hallucinerer".

Hallucination

GenAI-indhold, som virker troværdigt, men som er faktisk forkert.

Når en LLM skal vælge hvilken token den skal generere, vil der ofte være flere gode muligheder at vælge mellem, for eksempel fordi teksten kan anvende synonyme ord, fordi teksten kan præsentere pointer i forskellig rækkefølge, eller fordi teksten kan fortsætte i indholdsmæssigt forskellige retninger. I mange tilfælde er det en fordel, at en GenAI genererer tekst med et vist element af tilfældighed, fordi den derved kan opnå større sproglig fleksibilitet og variation. I stedet for at modellen altid genererer præcis den token, som er den bedst egnede, kan den vælge tilfældigt mellem de bedste tokens. GenAI-modeller har derfor normalt en indstilling, der styrer hvor meget tilfældighed modellen skal introducere – et såkaldt hyperparameter for "temperatur". Hvis temperaturen skrues højt op, kan to indtastninger af det samme prompt føre til skabelse af meget forskelligt indhold. Hvis temperaturen omvendt sættes lavt, så vil det samme prompt altid give indhold som er næsten eller komplet identisk. Kort sagt: jo højere temperaturen er indstillet, jo mere kreativ og varieret er GenAI, når den skaber indhold.

Kombinationen af, at GenAI reproducerer sproglige mønstre, som den ikke forstår, og at GenAI i en vis grad generer indholdet tilfældigt, har fået kritikere til at karakterisere LLM'er som "stokastiske papegøjer".¹²

Fintuning og indholdshegn

En LLM er en statistisk repræsentation af menneskeligt sprog. Det er på engang modellernes



¹² Se Bender, et al. 2021. Men se også Arkoudas 2023.

styrke og svagheit, at de ikke er designet til at løse bestemte opgaver, men kan være grundlag for løsningen af mange forskellige typer opgaver. Ofte kræver løsning af en mere konkret og specialiseret opgave imidlertid, at modellen justeres, ved at den trænes på mere specialiserede data om netop denne type opgave. Det kaldes "fintuning".

Fintuning

Yderligere træning af en allerede udviklet LLM, for at tilpasse modellen til en bestemt anvendelse.

Fintuning foregår ved, at udvikleren laver et datasæt med eksempler på den type opgave, som modellen skal specialiseres i at løse. Hvis eksempelvis en myndighed ønsker at udvikle en chatbot, kan dette være spørgsmål, som borgere plejer at stille, samt de svar som menneskelige medarbejdere har givet på sådanne spørgsmål.

Den kunstige intelligens gennemgår derpå en særlig træningsproces, så modellen bliver tilbøjelig til at efterligne træningseksemplerne, når den genererer indhold.

En særlig motivation for at fintune en LLM er, at man i den forbindelse kan bygge såkaldte "indholdshegn" (eng. "guardrails") ind i modellen.¹³ En LLM er i udgangspunktet i stand til at generere alle typer indhold. Det betyder, at det er muligt for en model at generere indhold som i nogle sammenhænge vil være upassende eller skadeligt (se kapitel 2). Modeller kan eksempelvis generere racistisk hadtale eller sexistisk chikane, og grafisk seksuelt eller voldeligt indhold. I nogle tilfælde kan modeller generere sådant indhold på baggrund af en uskyldig prompt, men i mange tilfælde vil det også være upassende, at GenAI er i stand til at generere sådant indhold, når brugeren ønsker det.¹⁴ Eksempelvis vil det være indlysende uhensigtsmæssigt, hvis brugere kan prompte en chatbot fra en offentlig myndighed til at generere racistisk eller seksuelt indhold, uanset om brugeren måtte ønske dette.

Indholdshegn er tekniske begrænsninger på, hvilke typer indhold en GenAI kan generere. En almindelig måde at bygge indholdsrestriktioner ind i en LLM er ved såkaldt "Reinforcement-Learning from

Indholdshegn

Tekniske begrænsninger i en GenAI, som reducerer eller forhindrer generation af uønsket indhold.

¹³ Indholdshegn kan bygges på flere måder. Det er almindeligt at kombinere fintuning med såkaldt "prompt-engineering", hvor udvikleren automatisk tilføjer skjulte instrukser til alle prompts fra brugeren, eller indholdsfiltere, som kan analysere og afvise både prompts og det genererede indhold. Se Ayyamperumal & Ge 2024; Dong, et al. 2024; Conitzer & Leben 2024.

¹⁴ Når dette er tilfældet i mange, snarere end i alle tilfælde, er det fordi der kan være undtagelser, for eksempel hvor en bruger ønsker at generere upassende indhold til træningsscenarier, eller hvor ansvarlige voksne ønsker at generere grafisk seksuelt eller voldeligt indhold til privat brug.

Human Feedback” (RLHF). RLHF er en form for fintuning, hvor udvikleren skaber en database med mærkede eksempler på uønsket indhold, for eksempel racistisk hadtale, og træner modellen til at undgå generation af indhold, der minder om disse eksempler. Træningen er ofte iterativ, således at modellen testes ved at prompte den på måder, som kunne lede den til at generere uønsket indhold. Resultaterne af denne test kan vurderes af mennesker, og eksempler som fortsat indeholder uønsket indhold kan mærkes, hvorpå modellen kan finjusteres igen på de nye eksempler. Gennem denne træningsproces justeres modellen gradvist, på en måde som forhindrer den i at generere de uønskede typer indhold.

Mange af de store, kommercielle GenAI-systemer er allerede fintunet, når almindelige brugere møder dem, for eksempel for at lære generelle adfærdsmønstre for chatbots, og for at bygge indholdshegn, som forhindrer generation af upassende og skadeligt indhold.

RAG-modeller

En LLM er normalt trænet på store datasæt, med et bredt udvalg af forskellige tekster. Det giver modellen en grundlæggende forståelse af menneskeligt sprog, men betyder, at modellen kan mangle specialiserede data, som er relevante for en særlig kontekst, og at den information, modellen har adgang til, kan være forældet. Hvis man vil anvende GenAI indenfor et specifikt domæne, kan det ofte være en fordel, at modellen har adgang til specialiseret information om netop dette domæne, inklusive information, som er helt opdateret. For eksempel kan en myndighed, der ønsker at bruge GenAI i en chatbot til medarbejdere, have brug for, at modellen har adgang til specifik og aktuel viden om myndighedens arbejde, procedurer og regler. Det kan imidlertid være dyrt og tidskrævende løbende at fintune modellen, så den har adgang til disse data.

En udbredt løsning på denne udfordring er at bruge såkaldte RAG-modeller (eng. "Retrieval-Augmented Generation"). En RAG-model kan tilgå en database med relevante informationer, som udvikleren enten selv har skabt eller har sikret adgang til. Når modellen modtager en prompt, søger den i databasen for at identificere information, som er relevant

RAG-model

En GenAI model, som kan trække på en database med relevant, specialiseret viden, når den skal generere indhold.

for denne prompt. De relevante informationer føjes til brugerens prompt. Derpå genererer modellen tekst med udgangspunkt i denne udvidede prompt.¹⁵

Ved at tilføje relevant information til brugerens prompt, får modellen mulighed for at generere tekst med udgangspunkt i domænespecifik og opdateret information. Ved at tilføje relevant information, reducerer man også modellens risiko for at generere tekst med misvisende eller forkert information (se kapitel 2). RAG har endvidere den fordel i sammenligning med fintuning, at det typisk er væsentligt hurtigere og billigere, at udvikle en RAG-løsning, end et tilsvarende fintunet system.



¹⁵ Gao, et al. 2023; Gupta, et al. 2024.

2. Generelle dataetiske udfordringer ved generativ AI

Generativ AI (GenAI) er en kraftfuld ny digital teknologi, som har potentiale til at forbedre mange processer i både den private og offentlige sektor, og skabe nye produkter for kunder og borgere. Men som med andre nye teknologier introducerer udvikling og anvendelse af GenAI både muligheder og risici. Forskere, aktivister og tænketanke har peget på en række dataetiske udfordringer.¹⁶ De kan løseligt klassificeres som udfordringer knyttet til almindelig anvendelse af GenAI, udfordringer knyttet til potentialet for misbrug af GenAI, og udfordringer knyttet til træning og bredere effekter af GenAI.

Generelle dataetiske udfordringer i udvikling og anvendelse af GenAI

Udfordringer i anvendelsen af GenAI	Udfordringer med potentiale for misbrug af GenAI	Udfordringer ved træning og bredere effekter af GenAI
Fejl og hallucinationer	Misinformation	Brud på intellektuel ejendomsret
Algoritmisk bias	Deep fakes	Miljøpåvirkning
Læk af private data	Manipulation	Udbytende arbejdspraksisser
Diffus ansvarlighed	Spredning af farlig information	Forstærkede tech-oligopoler
Uigennemsigthed	Forklædning af AI-genereret indhold	Forskydninger på arbejdsmarkedet
		Eksistentielle risici ved AGI
		Mistillid

¹⁶ For overblik, se Janjeva, et al. 2023; Myrstad 2023; Weidinger, et al. 2021; Bommasani, et al. 2021; Lorenz, et al. 2023; Bender, et al. 2021; Marchal, et al. 2024; Fang & Perkins 2024.

2.1. Udfordringer i anvendelse af generativ AI

Fejl og hallucinationer: GenAI er ikke i stand til at evaluere, om information i de tekster og billeder som den kunstige intelligens skaber eller bruger som input, er sand eller falsk. Det betyder, at GenAI kan generere indhold, som er faktisk forkert eller på andre måder fejlbehæftet (se kapitel 5). Ind imellem synes GenAI til tekstgeneration endda at fabulere, idet den skaber komplekse, detaljerede, og umiddelbart troværdige tekster om faktuelle forhold, som ingen basis har i virkeligheden. Det kaldes, at den kunstige intelligens "hallucinerer". Fejl og hallucinationer kan forårsage skade, når brugere vildledes, især hvis informationen angår vigtige spørgsmål, f.eks. personlige finanser eller sundhed.

Algoritmisk bias: Ligesom andre former for kunstig intelligens har GenAI risiko for at reproducere eller skabe bias, dvs. systematiske forskelle i den måde, GenAI behandler relevante grupper.¹⁷ GenAI som trænes på data med bias vil eksempelvis have tendens til at reproducere disse bias. Bias i træningsdata kan for eksempel være hadtale i de tekster, som en LLM trænes på. Det kan også være stereotyper i billeder, som forbinde visse professioner eller måder at se ud på med et bestemt køn eller en bestemt etnicitet. GenAI kan også fungere forskelligt for forskellige grupper, for eksempel ved at producere indhold af højere kvalitet for brugere, hvis data er overrepræsenteret i træningen, og lavere kvalitet indhold for brugere, der ikke ligner de overrepræsenterede grupper.

Læk af private data: En anden alvorlig udfordring ved anvendelsen af GenAI er, at sådanne systemer har en risiko for at lække træningsdata.¹⁸ GenAI lærer at skabe indhold ved at finde sammenhænge i træningsdata. Disse sammenhænge kan være så stærke og detaljerede, at GenAI kan reproducere elementer af træningsdata, når det skaber indhold. Dette inkluderer både data fra det oprindelige træningssæt, og data indtastet i en GenAI-applikation af brugere, når disse data lagres og bruges til yderligere træning af modellen. Sådanne læk af private data kan ske tilfældigt, men forskere og aktivister har også demonstreret, at det er muligt at "angribe" et GenAI-system, for at narre det til at reproducere træningsdata.¹⁹ Det foregår i prak-

¹⁷ Mange af de mest prominente eksempler på algoritmisk bias i GenAI er stereotyper i billedgeneration. Se Bianchi, et al. 2023; Nicoletti & Bass 2023. For et overblik over bias i GenAI med fokus på tekst-generation, se Gallegos, et al. 2024. For et studie af kønnet bias i tekstgeneration, se Kapoor & Narayanan 2023.

¹⁸ Yao, et al. 2024; Yan, et al. 2024.

¹⁹ Nasr, et al. 2023.

sis ved, at brugeren anvender usædvanlige prompts, som er designet til at skabe denne effekt. I nogle tilfælde er det lykkedes forskere, aktivister og journalister, at få GenAI til at reproducere følsomme data, som modellen er trænet på, næsten ordret.

Diffus ansvarlighed: Indhold skabt af GenAI er resultatet af en delvist tilfældig proces, som inkluderer brugerens input. Ansvar for konsekvenserne af GenAI kan være vanskeligt at placere mellem for eksempel udvikleren af en LLM, udviklerens underleverandører, som har samlet, organiseret og kodet træningsdata, udviklere som har fintunet eller implementeret modellen i en lokal kontekst, og slutbrugeren.²⁰ Det betyder både at det kan være svært at sige, hvem som er moralsk ansvarlige, for de udfordringer en GenAI model rejser, for eksempel hvis den begår fejl, og at der kan opstå en uheldig incitamentstruktur, hvor ingen føler sig forpligtede til at forebygge fejl eller skader.

Uigennemsigthed: GenAI-modeller er ekstremt komplekse, i en grad så det er umuligt, selv for de udviklere, som arbejder med at lave modellerne, at forstå, præcis hvordan et givent prompt bliver til en bestemt form for genereret indhold.²¹ Udviklere har også stærke incitamenter til at holde elementer af både træningsdata, modeller, træningsprocesser og kode hemmelige for at beskytte forretningsinteresser, og for at forhindre ondsindede aktører i at udnytte svagheder. Det betyder at GenAI i vid udstrækning er uigennemsigtig – vi kan ikke forstå hvordan den kunstige intelligens fungerer af både tekniske grunde, og når udviklere afviser at give indsigt i modellerne.

2.2. Udfordringer med potentiale for misbrug af generativ AI

Misinformation: En af de mest åbenlyse udfordringer ved GenAI er, at den kan misbruges til at skabe eller sprede falsk og skadelig information.²² GenAI kan eksempelvis generere phishing- eller svindeltekster, hadtale, konspirationsteorier og propaganda. Et prominent eksempel er brugen af GenAI til at skabe falske personaer – tilsyneladende rigtige personer, der ikke eksisterer – som kan udgive sig for at være egentlige personer, og støtte spredningen af misinformation, for eksempel ved at dele misinformation på sociale medier. GenAI personaer kan fremstå langt mere overbevisende

²⁰ Novelli, et al. 2024.

²¹ Zhao, et al. 2024.

²² Se Xu, et al. 2023.

end klassiske bots, fordi GenAI gør en chatbot i stand til at engagere sig med modtagere, for eksempel for at besvare spørgsmål, og understøtte falske påstande med argumenter. GenAI misinformation kan således gøre det lettere for svindlere at narre modtagere til at falde for finansielle svindelnumre, eller for påvirkningsagenter at skabe indtrykket af, at andre personer tror på eller støtter falske eller ekstreme påstande.²³

Deep fakes: Misinformation kan også misrepræsentere specifikke, virkelige personer ved at skabe indhold med virkelige personer i fiktive situationer, som i de kendte billeder af paven i en stilfuld dynejakke eller Donald Trump, som bliver arresteret.²⁴ Sådant misinformation kan skade den person som fremstilles, hvis informationen er ydmygende eller inkriminerende. Et særligt problematisk eksempel er såkaldt deep fake pornografi, som placerer virkelige personer – i praksis næsten altid kvinder – i falsk pornografisk materiale.²⁵ Denne misinformation kan have til formål at skade den virkelige person, som fremstilles i en fiktiv situation, men det kan også udnyttes til at støtte falske påstande, ved at placere personer som er autoriteter, i en situation hvor de støtter påstanden, for eksempel for at skabe malplaceret tillid til svindelnumre.

Manipulation: GenAI kan også bruges til at skabe manipulerende indhold, som er skræddersyet den individuelle modtagers psykologiske og kognitive svagheder.²⁶ I sådanne tilfælde kan indholdet være misinformation, men det kan også være information som er faktisk korrekt, men omhyggeligt udvalgt for at skubbe modtagerens holdning i en bestemt retning.

Spredning af farlig information: GenAI kan også risikere at generere faktisk korrekt men farlig information. Det kunne for eksempel være detaljerede anvisninger om hvordan man stjæler en bil, eller hvordan en teenager kan udøve selvskade. I sådanne tilfælde forårsager GenAI skade ved at gøre faktisk korrekt men skadelig information lettere tilgængelig.²⁷

Forklædning af AI-genereret indhold: En sidste udfordring med misbrug er, at AI-genereret indhold kan fremstilles som indhold skabt af rigtige personer. GenAI kan skabe indhold, som brugeren udgiver for at være sit eget, som i den omdiskuterede udfordring med, at elever og studerende afleverer opgaver med tekst, som er genereret

23 Se Europol 2023; Buchanan, et al. 2021.

24 Se Nguyen, et al. 2022; Naitali, et al. 2023; UK 2024.

25 Flynn, et al. 2021.

26 Simchon, et al. 2024; Matz, et al. 2024.

27 Se Europol 2023; Kapoor, et al. 2024; Bostrom 2011.

af GenAI (GenAI-plagiat). Men der kan også være tale om indhold, som brugeren udgiver for at være skabt af andre, som de kendte eksempler på brugere, der har skabt musiknumre med syntetiske udgaver af berømte sangers stemmer. Forklædning af indhold skabt af GenAI kan udgøre et problem, selv hvis selve indholdet ikke er faktisk forkert eller på anden vis skadeligt, idet forklædningen lader brugeren lukrere på, at misinformere andre om indholdets oprindelse.

2.3. Udfordringer ved bredere effekter af generativ AI

Brud på intellektuel ejendomsret: GenAI kan også rejse dataetiske udfordringer knyttet til processen med at udvikle LLM, og de bredere effekter, som anvendelse af GenAI kan have. En første sådan udfordring drejer sig om intellektuel ejendomsret for de data, som LLM'er trænes på. En LLM trænes på enorme datasæt, som høstes fra offentligt tilgængelige dele af internettet. Kritikere har påpeget, at mange af disse data kan være ophavsretsligt beskyttet, og i den forbindelse fremført, at træningen udgør en moralsk problematisk eller ulovlig udnyttelse af disse data.²⁸

Miljøpåvirkning: Udvikling og anvendelse af GenAI er også ressourceintensivt, især hvad angår energi og vand (til køling af datacentre). Kritikere har rejst bekymringer over, at GenAI vil forbruge energi i et omfang, der negativt påvirker den globale klimakrise, og kan gøre sparsomme vandressourcer vanskeligt tilgængelige i sårbare samfund.²⁹

Udbytende arbejdspraksisser: Finjustering af en LLM, for eksempel for at indbygge indholdshegn, kræver at modellen har adgang til store, omhyggeligt og manuelt mærkede datasæt. Disse datasæt fokuserer ofte på eksempler på indhold, som udvikleren ønsker at undgå eller begrænse, for eksempel hadtale, seksuelt indhold, og grafiske beskrivelser af vold. Kritikere har påpeget, at udviklere i mange tilfælde har placeret arbejdsintensive processer i lande, hvor arbejdskraften er billig, svagt organiseret, og savner effektiv juridisk beskyttelse. I nogle tilfælde, hævder kritikere, udnytter disse udviklere prekære arbejdsforhold i sårbare situationer, ved at tilbyde ringe arbejdsvilkår og lav løn for at udføre det psykologisk hårde arbejde med at mærke datasæt med uønsket indhold.

²⁸ Chesterman 2024; Lucchi 2023.

²⁹ Se Patterson, et al. 2021; Li, et al. 2023.

Forstærkede tech-oligopoler: GenAI er i dag et produkt på et marked, hvor det kræver enorme investeringer, at udvikle de mest avancerede former for kunstig intelligens. Det betyder, at det i udgangspunktet er meget få aktører, som har mulighed for at være med til at udvikle avanceret GenAI. Samtidig er der potentiale for, at de produkter, som først kommer ind på markedet, kan etablere særstatus på markedet, ved at definere standarder og indsamle brugerdata. Givet det enorme økonomiske potentiale ved GenAI kan monopoler eller oligopoler være både ekstremt profitabelt, og placere de(t) vindende firma(er) i en magtfuld position, til at påvirke tech-markedet, samt sociale og økonomiske forhold i bredere forstand.

Forskydninger på arbejdsmarkedet: GenAI vil øge effektiviteten af meget menneskeligt arbejde. Det er sandsynligt, at GenAI i nogle brancher vil gøre medarbejdere overflødige – enten fordi færre menneskelige arbejdere er nødvendige for at skabe samme resultat, eller fordi jobfunktioner kan overtages helt af AI. Effekten vil variere på tværs af sektorer og jobfunktioner, og effektivitetsgevinster vil i nogle tilfælde betyde, at arbejdere kan lægge mere indsats i de tilbageværende opgaver. Men i andre tilfælde kan store grupper arbejdere på kort tid oplever, at deres stillinger bliver nedlagt. GenAI risikerer altså at skabe massive forskydninger på arbejdsmarkedet. Det kan gøre det nødvendigt at tage politiske initiativer, for eksempel til at omskole de berørte personer, for at undgå at store grupper pludselig står uden arbejde, og den sociale uro som dette kan skabe.³⁰

Eksistentielle risici ved AGI: GenAI er den mest avancerede AI, mennesker hidtil har skabt. GenAI har også ifølge nogle forskere vist såkaldt emergente evner, det vil sige kvalitativt nye egenskaber for problemløsning, som opstår uventet, som et resultat af skalering af andre egenskaber.³¹ Emergente evner kan potentielt gøre en kunstig intelligens mere sofistikeret og fleksibel. Klassisk kunstig intelligens er højt specialiseret – den er designet til at løse en snævert defineret opgave, og kan ikke løse andre opgaver end netop denne. En kunstig intelligens, som er i stand til at løse en lang række forskellige problemer på niveau med mennesker kaldes for en AGI – en generel kunstig intelligens (eng. "Artificial General Intelligence"). Prominente forskere har argumenteret for, at LLM'er er farligt tæt på en AGI, som kan igangsætte en iterativ og accelererende proces med at forbedre sig selv. En sådan proces, hævder nogle forskere, kan meget hurtigt udvikle en kunstig superintelligens, som kan udgøre en eksistentiel risiko for mennesker.³²

30 Se Webb 2019; Eloundou, et al. 2023; Felten, et al. 2023; Santoni de Sio 2024.

31 Se Bubeck, et al. 2023.

32 Se Bostrom 2014.

Mistillid: Udover de konkrete skadevirkninger, som kan opstå som et resultat af skabelsen af forskellige typer indhold, har nogle debattører udtrykt bekymring for, at GenAI kan have en mere generel og diffus effekt i form af øget mistillid til korrekte informationer. En sådan effekt kan enten være et resultat af en overflod af falsk information, eller af bevidstheden om, at der findes lettilgængelige GenAI værktøjer, som ville kunne producere lignende indhold. Eksempelvis kunne nogle mistænke, at en lydfil eller et foto, som afslører en korrump politiker, er indhold skabt af GenAI, og derfor afvise at tage stilling til, om politikeren er korrump.

2.4. Hvilke dataetiske udfordringer er særligt relevante for offentlige myndigheders anvendelse af generativ AI?

Som den ovenstående gennemgang har vist, kan GenAI rejse en lang række dataetiske udfordringer, men det er ikke givet, at alle udfordringer er lige relevante i en bestemt kontekst. Offentlige myndigheder er en særlig aktør, som anvender GenAI på en særlig måde, og med et særligt ansvar. Hvilke dataetiske udfordringer er det relevant for offentlige myndigheder at forholde sig til, når de ønsker at anvende GenAI?

En række udfordringer knytter sig til potentialet for misbrug af GenAI, for eksempel skabelse af deep fakes, eller brug af GenAI til at målrette internetsvindler mod individuelle ofre. Disse udfordringer gør det relevant at overveje, for eksempel om man bør regulere hvilke typer GenAI, som er offentligt tilgængelige, eller stille krav om deklaration og såkaldt "vandmærkning" af indhold genereret med GenAI. Udfordringer som drejer sig om potentielt misbrug er imidlertid i begrænset omfang relevante for offentlige myndigheders anvendelse af GenAI, idet man må forudsætte, at myndigheder ikke misbruger GenAI til sådanne formål.

En anden gruppe udfordringer knytter sig til de bredere effekter af GenAI. Det drejer sig eksempelvis om hvordan udvikling og anvendelse af GenAI kan belaste miljøet, skabe forskydninger på arbejdsmarkedet, eller krænke intellektuel ejendomsret. Disse udfordringer kan være relevante for offentlige myndigheder. Eksempelvis kunne det have betydning for en myndighed, hvis den anvender en LLM som trækkes af markedet, fordi modellen har vist sig at være ulovligt trænet på ophavsretligt beskyttet data. Men det principielle spørgsmål om, hvilke data modeller etisk set bør trænes på, er i mere begrænset omfang relevant for den enkelte myndighed, idet der fortrinsvis er tale om et spørgsmål, som må behandles på politisk niveau. Tilsvarende vil udfor-

dringer om miljøpåvirkning og forskydninger på arbejdsmarkedet i nogen grad være udfordringer, som det er mere relevant for politiske beslutningstagere at forholde sig til, for eksempel ved at beslutte hvordan den offentlige sektor som hele skal agere, snarere end udfordringer, som den enkelte myndighed kan løse.

Når en offentlig myndighed overvejer at anvende GenAI, så er de mest umiddelbart relevante udfordringer derfor måske fortrinsvis de udfordringer, som knytter sig til selve anvendelsen af GenAI, herunder risiko for fejl, bias, og læk af private data. Det er den måske mest grundlæggende af disse udfordringer, risikoen for fejl, som er fokus for Dataetisk Råds arbejde i denne rapport.



3. Generativ AI ved danske offentlige myndigheder

Generativ AI (GenAI) er en kraftfuld ny teknologi, og den offentlige sektor i Danmark er måske verdens mest digitaliserede. Det er således ikke overraskende, at offentlige myndigheder har vist stor interesse for den nye type kunstig intelligens, og at man mange steder er gået i gang med at eksperimentere med teknologien.

De dataetiske udfordringer, som GenAI kan rejse, afhænger ofte af præcis hvordan teknologien udvikles og anvendes. Derfor er det vigtigt, når man vil analysere de dataetiske udfordringer, at gøre sig klart, hvordan teknologien aktuelt anvendes, og hvilke brugsscenarier som er realistiske indenfor en rimelig tidshorisont. Omvendt bør man bestræbe sig på at undgå at generalisere udfordringer fra rent hypotetiske og urealistiske scenarier.

Dette kapitel giver et overblik over hvordan GenAI anvendes i den offentlige sektor i Danmark, herunder hvordan forskellige måder at anvende teknologien kan variere på måder der gør en dataetisk forskel.

3.1. Kilder til Dataetisk Råds overblik

Det er vanskeligt at opnå et samlet overblik over hvordan offentlige myndigheder i Danmark anvender GenAI, eller kunstig intelligens i bredere forstand, fordi det ikke er lovpligtigt for offentlige myndigheder at oplyse om dette, ligesom der ikke findes en myndighed med ansvar for at skabe og løbende opdatere en sådan oversigt.

Ikke desto mindre findes der en række kilder, som kan hjælpe med at give et i hvert fald delvist dækkende overblik.

Version2, et dansk online-medie med særligt fokus på digital teknologi, udførte i sommeren 2024 en rundspørge til samtlige danske kommuner. Den resulterende database indeholder information

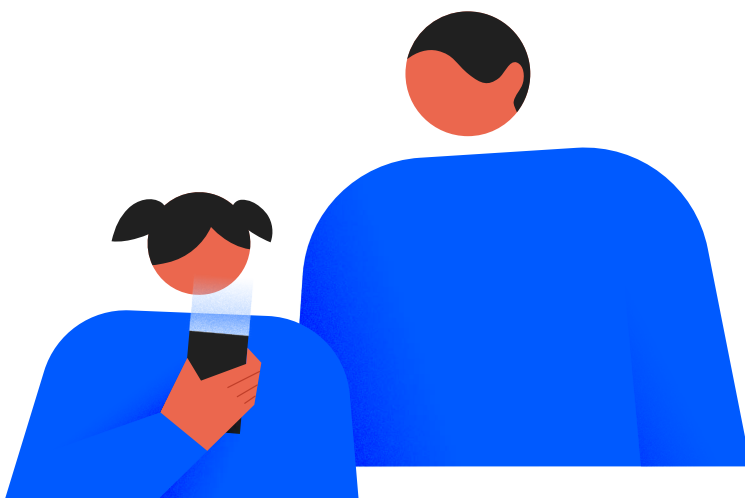
GenAI i den kommunale sektor

Dataetisk Råd vurderer, at godt 2/3 af danske kommuner arbejdede med GenAI i en eller anden udstrækning i 2024.

om hvorvidt kommunen anvender GenAI, herunder hvilken type (for eksempel Co-Pilot vs ChatGPT), samt retningslinjer for brug, men ikke antal og type af projekter i den enkelte kommune. I visse tilfælde er de svar, Version2 har fået, uklare, men for en rimelig fortolkning viser rundspørgen, at 69 ud af de 98 kommuner per sommeren 2024 arbejdede med GenAI i en eller anden form – i nogle tilfælde i form af løsninger, som var i drift, men i mange tilfælde kun i en afsøgende eller testfase.

Kommunernes Landsforening har oprettet en database over projekter med kunstig intelligens i den kommunale sektor, det såkaldte AI landkort.³³ Landkortet giver en kort præsentation af hvert af de inkluderede projekter, samt kontaktoplysninger til personer knyttet til projektet. Det er frivilligt for den enkelte kommune at indmelde projekter til databasen, hvorfor det er muligt at der findes GenAI-projekter, som ikke optræder i databasen. Databasen er designet til at rumme både GenAI-projekter og projekter med andre former for kunstig intelligens, men ved manuel gennemgang af basen er det muligt at identificere en række projekter med GenAI.

Forskningsprojektet AI & Human Rights, ved Jura på Københavns Universitet har oprettet en database over offentlige myndigheders brug af kunstig intelligens.³⁴ Databasen er baseret på løbende overvågning af projekter i den offentlige sektor, og inkluderer således både kommunale, regionale og statslige projekter. Når et projekt identificeres, udarbejder projektet et standardiseret oplysningsskema på baggrund af en aktindsigtsanmodning. Skemaet indeholder information om projektets formål, tekniske karakter og risikoprofil. Databasen er ikke offentlig tilgængelig, men medarbejdere ved forskningsprojektet har bistået Dataetisk Råd med at identificere offentlige aktører, som anvender GenAI.



33 <https://www.kl.dk/videncenter/viden-og-vaerktoejer/ai/kommunernes-ai-landkort>

34 <https://jura.ku.dk/icourts/research/artificial-intelligence-and-human-rights/>

Digitaliseringsstyrelsen udarbejdede i 2024 et inspirationskatalog, som præsenterer syv udvalgte offentlige projekter med GenAI, baseret på styrelsens egen kortlægning af hvor og hvordan GenAI anvendes i den offentlige sektor.³⁵

Dataetisk Råd har også været i dialog med fem udvalgte offentlige aktører, som arbejder med GenAI:

- ATP
- Gladsaxe kommune
- Miljøstyrelsen
- Vejle kommune, og
- Århus kommune

De fem aktører blev udvalgt fordi de har været hurtige til at udforske den nye teknologi, men repræsenterer forskellige typer myndighed, og bruger GenAI på forskellige måder. Dialogen har hjulpet rådet med at få et mere detaljeret og præcist indtryk af, hvordan den offentlige sektor anvender og kan anvende teknologien, gennem præsentationer af aktørernes projekter, samt de overvejelser de har gjort sig i forbindelse med arbejdet.

3.2. Hvad er det nye ved generativ AI?

Kunstig intelligens i bred forstand har som teknologi potentiale til både at udføre opgaver, som det ellers kræver menneskelig arbejdskraft at udføre, og til at forbedre opgaveløsning, enten ved at styrke menneskers evne til at løse opgaver, eller ved at udføre opgaver bedre end mennesker. En oplagt begrundelse for, at offentlige myndigheder udforsker eller anvender kunstig intelligens, er således den rimelige forventning om, at det i nogle tilfælde vil være muligt at frigøre menneskelige ressourcer, eller at forbedre kvaliteten i det offentlige løsnings af opgaver.

GenAI er en relativt ny og særlig form for kunstig intelligens, som kan bruges af offentlige myndigheder på mange forskellige måder. Flere offentlige myndigheder har hurtigt taget teknologien til sig. Den store interesse for GenAI skyldes dels, at GenAI

³⁵ https://digst.dk/media/31743/inspirationskatalog-syv-eksempler-med-generativ-ai-i-den-offentlige-sektor-webtilgaengelig_final.pdf

kan gøre det lettere at anvende kunstig intelligens, og dels at den kunstige intelligens kan anvendes på nye måder.

På et meget overordnet niveau er én væsentlig fordel ved GenAI, sammenlignet med mange andre former for kunstig intelligens, at GenAI kan håndtere såkaldt "ustrukturerede" data, og at den kan anvendes på tværs af forskellige beslutningsproblemer. Mange former for kunstig intelligens kræver strukturerede data både til træning og som input. Det betyder, at den kunstige intelligens skal bruge bestemte data, som er organiseret i en bestemt struktur.

Med et simpelt eksempel, så kunne en kunstig intelligens forsøge at gætte hvor høj en person er, ved at bruge data om alder, køn og vægt. I denne situation har en klassisk kunstig intelligens brug for at blive fodret med netop disse data, i et bestemt format, og en bestemt organisering. Den vil kunne bruge "34 år; kvinde; 65kg", men ikke "1990 (fødselsår); 0,065 tons; F", selvom de to serier data har de samme informationer, fordi data i den anden serie optræder i andre formater og en anden rækkefølge.

Behovet for at arbejde med strukturerede data kan være en alvorlig begrænsning ved kunstig intelligens, fordi det ofte er meget arbejdskrævende, at indsamle og rense data, så de får den struktur, den kunstige intelligens har brug for. Derfor kan det være en fordel, at GenAI kan behandle ustrukturerede data, for eksempel naturlig tekst, hvor ord og tegn kan optræde i vidt forskellige kombinationer og rækkefølge.

LLM'er har også en evne til at arbejde fleksibelt med naturligt sprog, som tidligere former for kunstig intelligens kun har haft i begrænset omfang. Det gør GenAI i stand til at arbejde med nye typer opgaver. Mange former for kunstig intelligens er designet til at løse en bestemt opgave, og kan ikke anvendes til at løse andre opgaver, selv nært beslægtede opgaver. Det betyder, at man skal træne en ny kunstig intelligens for hver ny opgave.

Den kunstige intelligens, som er trænet til at vurdere en persons højde, er specialiseret til at løse netop denne opgave, og vil typisk ikke kunne løse andre opgaver, ikke engang beslægtede opgaver, som at vurdere en persons skostørrelse. Men brugen af LLM giver GenAI en fleksibilitet, som ofte gør dem i stand til at løse mange forskellige opgaver indenfor et domæne. Eksempelvis vil en GenAI chatbot normalt være i stand til at besvare mange forskellige slags spørgsmål, snarere end en bestemt type spørgsmål.

3.3. Hvordan bruges generativ AI af offentlige myndigheder?

GenAI kan bruges af offentlige myndigheder på mange forskellige måder og til forskellige formål. Forskellige brugsscenarier kan rejse forskellige dataetiske udfordringer, eller den samme udfordring i henholdsvis større og mindre grad. Nogle brugsscenarier kan eksempelvis skabe risici for læk af følsomme data, mens andre kan skabe risici for fejlagtige beslutninger med betydelige konsekvenser for borgeren.

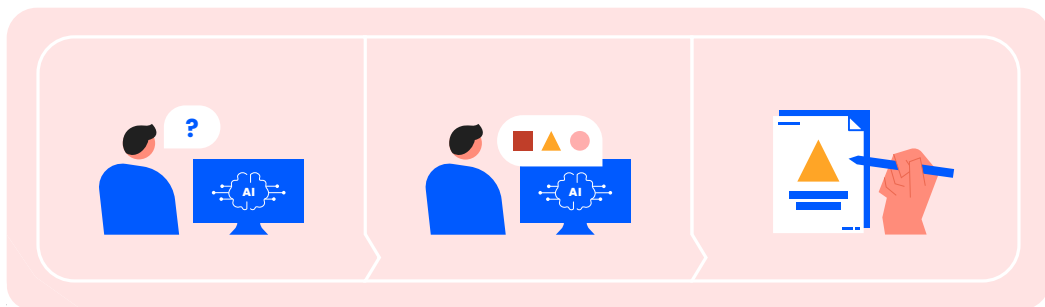
På et generelt niveau kan man skelne mellem seks forskellige brugsscenarier:

- Ide-generation
- Tekstredigering
- Udkast til brugerdefineret tekst
- Informationssøgning
- Opsummering af materiale
- Vurdering af sagsforhold

Brugsscenarierne kan variere yderligere afhængigt af for eksempel hvordan myndigheden integrerer GenAI i sit arbejde, og om brugeren er en medarbejder ved myndigheden eller en borger.³⁶

Ide-generation

En første almindelig måde, offentlige myndigheder kan bruge GenAI, er til ide-generation. GenAI kan på baggrund af en prompt komme med forslag til det, som brugeren spørger om. Det kan for eksempel være forskellige måder at kommunikere et budskab, eller organisere en begivenhed. I sådanne tilfælde fungerer GenAI groft sagt ligesom en menneskelig sparringspartner under en brainstorm-session.



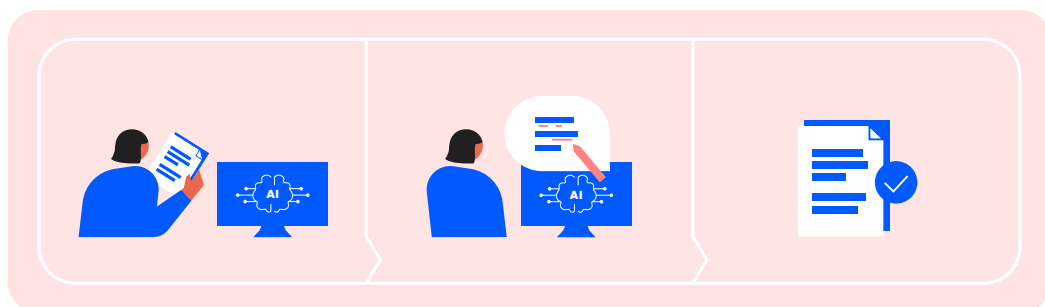
³⁶ For yderligere diskussion af forskellige typer brugsscenarier, se KL & Digitaliseringsstyrelsen 2024; Implement 2023.

Fordi der er tale om en kreativ proces, hvor GenAI genererer ideer til et menneske, som kan overveje og forkaste disse ideer, rejser idegeneration ofte relativt beskedne dataetiske udfordringer. Der kan være risiko for at lække private data, hvis brugeren inkluderer dem i en prompt. Denne risiko kan reduceres ved at træne medarbejdere til at skelne mellem følsomme og ikke-følsomme data, eller ved at bruge GenAI i et lukket miljø, hvor data i brugerprompt ikke deles med andre, herunder udvikleren af GenAI. Der kan også være risiko for bias i de ideer som GenAI producerer, for eksempel ved at de reproducerer sociale stereotyper. Til gengæld er det i begrænset omfang muligt for GenAI at generere fejlagtige svar, idet ideer typisk vil være for eksempel irrelevante, urealistiske eller uegnede snarere end faktisk forkerte. Af samme grund kan man formode, at der i mindre grad vil være behov for, at kunne placere et ansvar (ansvarlighed), eller forstå hvordan GenAI har genereret et svar (gennemsigtighed).

Offentlige myndigheder som ønsker at bruge GenAI til ide-generation, kan ofte anvende en ikke-specialiseret chatbot, som Google Gemini, Microsoft Copilot eller OpenAI ChatGPT.

Tekstredigering

GenAI kan også anvendes til tekstredigering. Mange GenAI applikationer er i stand til at rette for eksempel stavefejl, tegnsætning, og grammatik i en tekst, som brugeren deler. GenAI kan også foreslå synonymer, eller alternative måder at formulere en sætning. Det kan eksempelvis være et svar til en borger, eller et resumé af en sag, som en medarbejder ønsker at sikre, er klart og sprogligt korrekt. I sådanne tilfælde fungerer GenAI lidt ligesom en menneskelig redaktør der skriver kommentarer og foreslår korrektur.



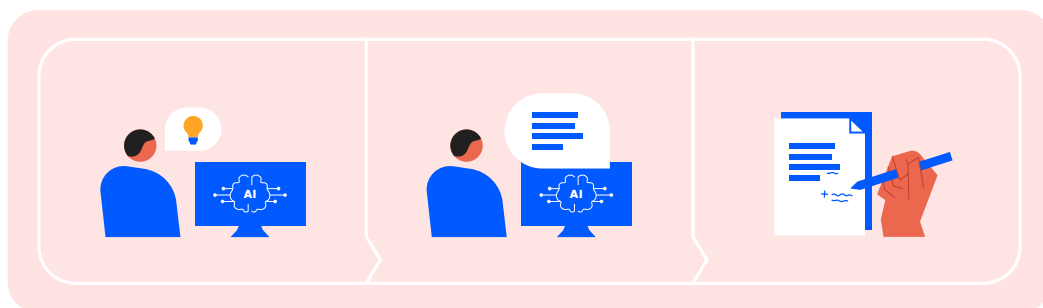
Fordi der er tale om en allerede forfattet tekst, hvor indholdet bestemmes af et menneske, som har mulighed for at anvende eller afvise de rettelser, som GenAI genererer, rejser dette brugsscenarie også ofte relativt beskedne dataetiske udfordringer. Der kan være risiko for at lække private data, hvis brugeren deler sådanne data som en del

af den tekst, GenAI skal rette. Ligesom for idegeneration kan denne risiko reduceres ved at træne medarbejdere til at skelne mellem følsomme og ikke-følsomme data, eller ved at bruge GenAI i et lukket miljø, hvor data i brugerprompt ikke deles med andre, herunder udvikleren af GenAI. Der kan også være risiko for fejl, i den forstand, at de foreslåede ændringer er sprogligt forkerte, eller ændrer vigtige betydninger i teksten. I begge tilfælde har den menneskelige forfatter mulighed for at undgå fejl, ved at undlade at anvende de foreslåede ændringer. Til gengæld vil der typisk i mindre grad være risiko for algoritmisk bias, ligesom der i begrænset omfang vil være behov for, at kunne placere et ansvar (ansvarlighed), eller forstå hvordan GenAI har genereret et svar (gennemsigtighed).

Ligesom for idegeneration kan offentlige myndigheder, som ønsker at bruge GenAI til tekstredigering, ofte anvende en ikke-specialiseret chatbot, som Google Gemini, Microsoft Copilot eller OpenAI ChatGPT.

Udkast til brugerdefineret tekst

GenAI kan også anvendes til at generere et egentligt udkast til en tekst, med et indhold som brugeren definerer. Brugeren kan eksempelvis definere en genre, en modtager, en afsender, en længde, et tonefald, centrale pointer, og en konklusion, hvorpå GenAI genererer et eller flere udkast til den pågældende tekst. Brugeren kan efterfølgende anvende, redigere i, eller afvise disse udkast. Det kan for eksempel være en medarbejder ved en myndighed, som beder om et udkast til et bestemt svar til en borger, en pressemeddelelse om et nyt initiativ, eller til en fremstilling af en sag for sine kollegaer.



Brug af GenAI til at skabe udkast til brugerdefineret tekst rejser potentielt dataetiske udfordringer. Fordi brugeren definerer indholdet i teksten, er det fortsat myndighedens medarbejder, som afgør hvordan teksten ser ud. Det betyder, at GenAI ikke laver automatiseret beslutningsstøtte i dette brugsscenarie – hvis medarbejdere beder GenAI om at skrive udkast til en tekst, uden at præcisere hvad indholdet skal være, så er der tale om et andet brugsscenarie, hvor GenAI vurderer sagsforhold (se nedenfor). Men

GenAI får større indflydelse på teksten, end når GenAI eksempelvis blot anvendes til redigering af en menneskeskabt tekst.

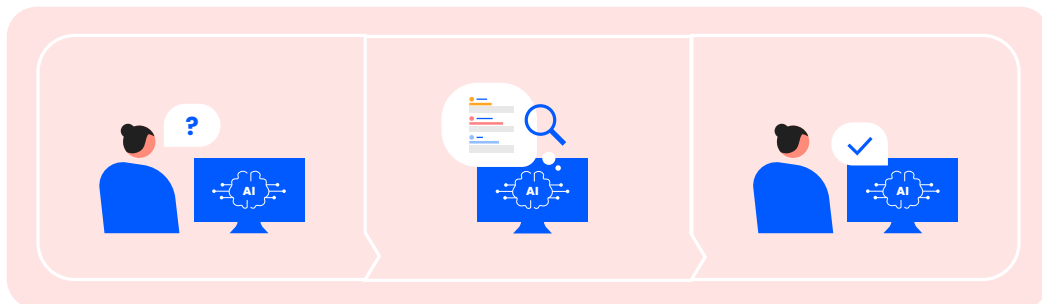
Ligesom ved andre brugsscenerier, kan der være en risiko for, at lække private data, hvis brugeren deler sådanne data som en del af den prompt til GenAI, om end denne risiko antageligt er mindre, end for eksempel når brugeren deler udkast til tekst, og kan modvirkes ved at træne medarbejdere, og arbejde i et lukket miljø. Der kan også være risiko for fejl, i den forstand at GenAI genererer tekst, som indeholder faktuelle fejl, eller som på forskellig vis ikke opfylder det formål, medarbejderen har med teksten. Et udkast til et brev, som skal informere en borger, kan for eksempel indeholde forkert information, udelade vigtig information, eller fremstille information på en måde, som gør det vanskeligt for borgeren, at forstå denne information. Det vil variere afhængigt af hvilken type tekst brugeren genererer, hvilke dataetiske konsekvenser en fejl vil have, og hvor alvorlige disse konsekvenser er. Det kan også spille en rolle, hvordan brugeren anvender udkastet, for eksempel hvor meget tid brugeren afsætter til at arbejde videre med teksten og identificere eventuelle fejl. Tekstgeneration øger også risikoen for algoritmisk bias, idet teksten kan variere på stereotyp vis i indhold, struktur eller tone afhængigt af den intenderede modtager, eller i selve indholdet kan reproducere stereotype eller stigmatiserende ideer eller udtryk. På grund af risikoen for sådanne fejl og bias, bliver det væsentligt i højere grad at kunne placere et ansvar for teksten. Men fordi tekstens indhold er defineret af mennesker kan man argumentere for, at det fortsat i mindre grad er afgørende, at man kan forklare hvordan GenAI har skabt indholdet.

Ligesom ved idegeneration og tekstredigering, vil det ofte være muligt for offentlige myndigheder, som ønsker at bruge GenAI til at skabe udkast til tekst, at anvende en ikke-specialiseret chatbot, som Google Gemini, Microsoft Copilot eller OpenAI ChatGPT.

Informationssøgning

Det fjerde brugsscenarie er GenAI til informationssøgning. Når GenAI anvendes på denne måde fungerer den som alternativ til andre måder at søge information, for eksempel med en generel søgemaskine i en browser, en søgefunktion på en bestemt hjemmeside, eller manuel søgning i en FAQ eller en vidensdatabase. Brugeren kan indtaste sin forespørgsel i naturligt sprog, og GenAI genererer et svar, som trækker på de data den kunstige intelligens har adgang til. Typisk vil GenAI til informationssøgning være en chatbot, ofte i en RAG-model, hvor GenAI har adgang til og trækker på et specialiseret datasæt, når den skal generere svar. Sådanne chatbots kan både være rettet mod medarbejdere ved en offentlig myndighed, som for eksempel kan søge informationer om regler, historiske sager og myndighedens politikker, og mod

borgere, som kan søge information af både mere praktisk og mere specifik karakter, fra åbningstider og kontaktmuligheder til regler og procedurer.



Brug af GenAI til informationssøgning kan rejse en række forskellige dataetiske udfordringer, som blandt andet afhænger af, hvorvidt GenAI anvendes af medarbejdere ved en myndighed eller borgere, som interagerer med myndigheden.

Når en GenAI til informationssøgning er rettet mod medarbejdere ved en myndighed, kan anvendelsen rejse mange af de samme udfordringer, som andre anvendelser. Der kan være risiko for at lække følsomme data, hvis medarbejdere prompter GenAI med disse informationer. Disse risici kan begrænses ved at træne medarbejdere, og arbejde i et sikkert miljø. Der er også en risiko for, at de svar GenAI giver, er fejlbehæftede. Hvis medarbejdere anvender fejlagtig information, når de sagsbehandler eller vejleder borgere, kan det både føre til procedurefejl i sagsbehandlingen, og til at der begås fejl, som skader borgere. Risikoen for sådanne fejl kan begrænses ved at træne medarbejdere til at forholde sig kritisk til den information, de får fra en GenAI søgemaskine, og ved at begrænse GenAIs kreativitet når den genererer svar. Ved at regulere "temperaturen" for et GenAI-system kan man for eksempel tvinge systemet til at generere svar, som er helt eller næsten identiske med informationstekst, som mennesker har produceret. En offentlig myndighed kan på denne måde få en GenAI chatbot til at generere svar, som direkte citerer eller let parafraserer relevant tekst fra myndighedens hjemmeside eller intranet. Ulempen ved på denne måde at begrænse GenAI, er for det første, at søgemaskinens svar i højere grad afhænger af den information, som mennesker har produceret. Denne information skal således være opdateret, fyldestgørende og korrekt. For det andet bliver GenAI i ringere grad i stand til at formulere svar, som er tilpasset den enkelte brugers specifikke forespørgsel og behov.

Når en søgemaskine rettes mod borgere forstærkes disse udfordringer. Borgere kan i mindre grad end medarbejdere ved en myndighed forventes at forstå risici ved at anvende en GenAI chatbot. Myndigheder kan heller ikke antage, at borgere har mu-

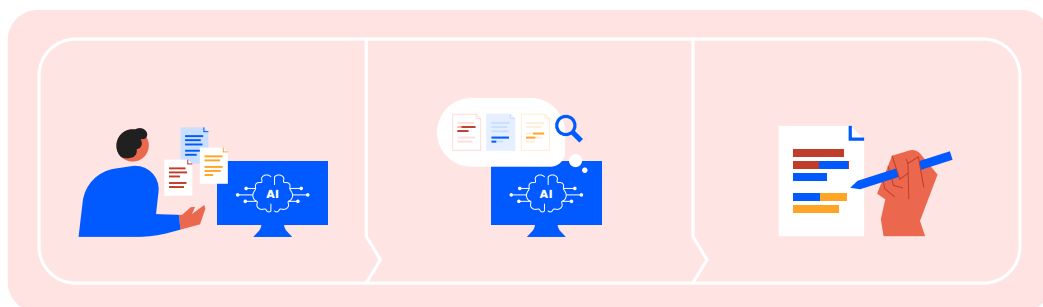
lighed for og kompetencer til kritisk at vurdere den information, som GenAI leverer. Udfordringernes størrelse og karakter afhænger fortsat i afgørende grad af hvilke informationer GenAI hjælper borgere med. Forkert information kan i nogle situationer blot være frustrerende for borgeren, men i andre situationer have alvorlige konsekvenser.

Uanset om GenAI anvendes til informationssøgning af borgere eller medarbejdere, kan det være væsentligt at gøre det klart, hvem som er ansvarlig for både generation af information, og den måde informationen bliver anvendt på. Omvendt vil der i begge tilfælde, fordi GenAI anvendes til at formidle faktuel information, være mere begrænsede udfordringer med uigennemsigthed.

Når en offentlig myndighed ønsker, at anvende GenAI til informationssøgning, vil det ofte være nødvendigt at udvikle en GenAI-løsning, for eksempel en chatbot, som er tilpasset det specifikke domæne. Det kan være en lokalt implementeret RAG-model, som får adgang til de relevante data, så den bliver bedre i stand til at svare på spørgsmål indenfor domænet, eller en model som fintunes på den relevante type forespørgsler, så den bliver bedre i stand til at generere velegnede svar.

Opsummering af materiale

GenAI kan også anvendes til at opsummere materiale. Brugeren kan prompte GenAI med det materiale, som skal opsummeres, instrukser om at generere et resumé af en given længde, og eventuelle særlige fokus på bestemte dele af materialet. Det kan eksempelvis være videnskabelige rapporter, sagsakter, eller en transskription af et møde, som brugeren ønsker et kort og letlæst resumé af, der gør det muligt på kort tid at få overblik over de væsentligste pointer.



Anvendelse af GenAI til opsummering af materiale har stort potentiale, men kan også rejse en række dataetiske udfordringer. Karakteren af disse udfordringer vil i nogen grad afhænge af, hvilket materiale det er, som bliver opsummeret, og hvordan resuméet anvendes, herunder hvilke muligheder medarbejdere eller borgere har for at vurdere,

om resuméet er fejlbehæftet. Anvendelsen rejser i alvorlig grad udfordringen med risiko for læk af følsomme data, i brugsscenarier hvor myndigheden ønsker at opsummere materiale, der indeholder følsom information. Det skyldes, at det er nødvendigt at dele dette materiale med GenAI, for at den kunstige intelligens kan lave et resumé af materialet. Af samme grund anvender offentlige myndigheder i de eksempler, som Dataetisk Råd er bekendt med, kun GenAI til opsummering af materiale i lukkede miljøer, hvor det teknisk kan garanteres, at de data brugeren deler med GenAI, ikke bliver tilgængelige udenfor myndigheden. Opsummering rejser også udfordringer med risiko for fejl og bias, som gør det vigtigt, at det er muligt at placere et ansvar for resuméets indhold, og den måde det anvendes på. Anvendelsen rejser risikoen for fejl, fordi resuméer skabt med GenAI kan indeholde forkert information, såkaldte hallucinationer, udelade relevant information, og præsentere information på uhensigtsmæssige måder, for eksempel ved at give resuméet en struktur der indikerer, at der bør lægges meget vægt på visse informationer, når disse informationer ret beset er mindre vigtige. Hvis sådanne fejl er ulige fordelt på tværs af relevante demografiske grupper, kan GenAI være under indflydelse af algoritmisk bias. Offentlige myndigheder kan bruge GenAI til at opsummere materiale på mange forskellige måder, for eksempel afhængigt af hvilket materiale myndigheden opsummerer, og hvordan den bruger det indhold, GenAI skaber. Nogle af variationer af brugsscenarier rejser væsentligt større dataetiske udfordringer end andre, men på et overordnet niveau betragter Dataetisk Råd brug af GenAI til opsummering af materiale som den dataetisk set mest udfordrende anvendelse af GenAI ved offentlige myndigheder i Danmark i dag.

Når GenAI anvendes til at opsummere materiale vil det ofte være en fordel at udvikle en specialiseret GenAI-løsning, for eksempel ved at fintune en model på eksempler, der består af både materiale og menneskeskabte resuméer. Derved forbedres den kunstige intelligens' evne til at skabe egnede resuméer.

Vurdering af sagsforhold

Det sidste brugsscenarie for offentlige myndigheder er anvendelse af GenAI til vurdering af sagsforhold. Meget af debatten om offentlige myndigheders brug af kunstig intelligens de senere år har fokuseret på automatiserede beslutninger, hvor den kunstige intelligens vurderer et sagsforhold.³⁷ Det kaldes også for "ADM" (eng. "automated decision-making"). Hvis en myndighed vil anvende GenAI til at vurdere et sagsforhold,

³⁷ Vi anvender her betegnelsen, at GenAI "vurderer et sagsforhold", snarere end det mere udbredte "automatiseret beslutningsstøtte", fordi visse andre brugsscenarier også kunne betegnes som varianter af automatiseret beslutningsstøtte, eksempelvis når en GenAI opsummerer materiale, og dette resumé lægges til grund for en myndigheds beslutning.

kan en medarbejder prompte den kunstige intelligens med information om sagen, og bede GenAI om at generere et udkast til en vurdering af det relevante forhold. Et eksempel på en sådan vurdering kunne være den endelige afgørelse i sagen. Denne vurdering kan enten gives til og indgå i overvejelserne hos en menneskelig beslutningstager – i så fald er der tale om automatiseret beslutningsstøtte – eller direkte omsættes i en afgørelse – i så fald er der tale om en fuldt automatiseret beslutning.

I dansk sammenhæng har der været debat om brug af klassisk, ikke-generativ kunstig intelligens til at vurdere sagsforhold blandt andet til tidlig opsporing af børn i mistrivsel, til vurdering af risiko for langtidsledighed, og til offentlige ejendomsvurderinger. Men også internationalt har ADM fået meget opmærksomhed, idet forskere og aktivister har udtrykt bekymring for de dataetiske udfordringer, som ADM kan rejse.

GenAIs mulighed for at arbejde med ustrukturerede data betyder, at det ofte ikke er nødvendigt at organisere de relevante data i et bestemt format, før den kunstige intelligens kan anvende dem. Det kan gøre det betydeligt lettere, at anvende GenAI til vurdering af komplekse forhold, end hvis man anvender klassisk kunstig intelligens. Til gengæld betyder LLM'ers store kompleksitet også, at det er uigennemsigtigt hvilke forhold den kunstige intelligens lægger vægt på, når den foretager sin vurdering. Selv med avancerede metoder til statistisk analyse er det kun i begrænset omfang muligt at forklare, hvordan GenAI generer et givet indhold (se kapitel 2).

Vurdering af sagsforhold er antageligt den mest kontroversielle måde, offentlige myndigheder kunne anvende GenAI. Når GenAI vurderer forhold i en sag, kan den kunstige intelligens få afgørende indflydelse på borgeres liv. Det er derfor vigtigt at gøre opmærksom på, at Dataetisk Råd ikke er bekendt med eksempler på, offentlige myndigheder i Danmark, som anvender GenAI til at vurdere sagsforhold eller automatiserede beslutninger. Af samme grund forholder rådet sig i denne sammenhæng ikke til de særskilte udfordringer, som offentlige myndigheders anvendelse af GenAI til vurdering af sagsforhold ville rejse. Der findes imidlertid enkelte eksempler fra andre lande, hvor offentlige myndigheder er gået i gang med at anvende GenAI til ADM.³⁸ Hvis myndigheder i Danmark beslutter at anvende GenAI til vurdering af sagsforhold, vil det i den forbindelse være relevant at overveje de særlige dataetiske udfordringer, som dette brugsscenarie rejser.

38 Se et nyligt eksempel fra den US amerikanske stat Nevada, hvor GenAI skal hjælpe med at vurdere ansøgninger om arbejdsløshedsstøtte: <https://gizmodo.com/googles-ai-will-help-decide-whether-unemployed-workers-get-benefits-2000496215>

4. Eksisterende dataetiske retningslinjer for offentlige myndigheders brug af generativ AI

Som vi har set i de forrige kapitler, kan offentlige myndigheders anvendelse af GenAI rejse en række forskellige dataetiske udfordringer. Flere danske og internationale aktører har udviklet retningslinjer for anvendelse af GenAI, som på forskellige måder forsøger at tackle disse dataetiske udfordringer. Det gælder blandt andet:

- KLS "Guide om offentligt tilgængelige tjenester med generativ AI".³⁹
- Digitaliseringsstyrelsens "Guide til offentlige myndigheder om ansvarlig anvendelse af generativ kunstig intelligens".⁴⁰
- Datatilsynets vejledning om "Offentlige myndigheders brug af kunstig intelligens".⁴¹
- Århus kommunes seks principper i rapporten "Kunstig intelligens fornyer velfærden".⁴²
- Det norske forbrugerråds rapport og anbefalinger "Ghost in the Machine".⁴³
- Det franske dataetiske råds (CNPEN) udtalelse "Systemès d'intelligence artificielle générative".⁴⁴

Nogle af disse retningslinjer og anbefalinger knytter sig til dataetiske udfordringer, som i mere begrænset omfang er relevante for offentlige myndigheders anvendelse af GenAI. Det gælder for eksempel anbefalinger om at indsætte såkaldte "vandmærker" i GenAI-indhold. Selvom det i særtilfælde kunne være relevant for offentlige myndigheder at indføre sådan vandmærkning, er der fortrinsvis tale om en løsning, som udviklere af LLM'er kan bruge til at forebygge, at ondsindede aktører genererer ind-

39 <https://www.kl.dk/media/h4anfadj/guide-om-generativ-ai-til-kommunerne.pdf>

40 <https://digst.dk/media/30849/110324-guide-til-offentlige-myndigheder-om-ansvarlig-anvendelse-af-generativ-kunstig-intelligens.pdf>

41 <https://www.datatilsynet.dk/Media/638321084132236143/Offentlige%20myndigheders%20brug%20af%20kunstig%20intelligens%20-%20Inden%20!%20g%C3%A5r%20!%20gang.pdf>

42 <https://faellesomaarhus.aarhus.dk/media/oyofqnys/kunstig-intelligens-forny-velfaerden-grafisk-version.pdf?format=nofor-mat>

43 <https://storage02.forbrukerradet.no/media/2023/06/generative-ai-rapport-2023.pdf>

44 https://www.ccne-ethique.fr/sites/default/files/2023-09/CNPEN_avis7_06_09_2023_web-rs2.pdf

hold, som kan præsenteres som autentisk eller menneskeskabt (for eksempel deep fakes eller GenAI plagiat).

Blandt de anbefalinger og retningslinjer, som knytter sig offentlige myndigheders anvendelse af GenAI, er der vist overlap mellem forskellige aktørers anbefalinger og retningslinjer. Det gælder især anbefalinger om:

- Deklaration af brug
- Menneskelig kvalitetssikring
- Beskyttelse af følsomme data i prompting
- Identifikation af bias
- Forsigtighed ved brug af automatiserede beslutninger

Deklaration af brug

En almindelig anbefaling er, at offentlige myndigheder tydeligt deklarerer når og hvordan de anvender GenAI, særligt i de situationer, hvor borgere på forskellig vis kommer i berøring med anvendelsen. Det kan for eksempel ske ved at myndigheden offentliggør denne information på en hjemmeside, men også ved at den enkelte borger gøres opmærksom på det, når borgeren har berøring med en myndigheds GenAI-systemer.

En udfordring for anbefalingen kan være, at det i nogle situationer kan forekomme utilstrækkeligt, alene at informere om at GenAI anvendes. Det kan imidlertid være vanskeligt at kommunikere mere detaljeret information om hvilket GenAI-system som anvendes, hvordan det virker, og hvilken rolle det spiller i myndighedens arbejde. I mange tilfælde vil sådan information være kompleks, og svær at formidle på en letforståelig måde, for en bred og varieret gruppe borgere.

Menneskelig kvalitetskontrol

Flere aktører anbefaler menneskelig kvalitetskontrol, ved at medarbejdere undersøger om der er fejl i GenAI-indhold. Det formuleres ofte sådan, at den kunstige intelligens bør støtte, men ikke erstatte medarbejders arbejde og beslutninger. Hvis GenAI for eksempel har lavet et udkast til et resumé af materiale, kan en menneskelig medarbejder gennemse teksten, og korrigere eventuelle fejl, inden resuméet anvendes på anden vis. Hvis GenAI anvendes til informationssøgning, kan medarbejderen verificere oplysninger ved at konsultere kollegaer eller andre kilder. Anbefalingen kan gøre det nødvendigt, at medarbejdere ved myndigheden trænes i at opdage de fejl, som karakteriserer GenAI-indhold, samt at medarbejdere får de nødvendige ressourcer, til at udføre effektiv kontrol med GenAI-indhold.

En udfordring for anbefalingen kan være, at der er visse typer fejl, som er lettest både at opdage og håndtere ved systematisk, statistisk analyse af et GenAI-system. Det kan også være vanskeligt at opnå effektiviseringsgevinster ved anvendelsen af GenAI, hvis myndigheden skal afsætte ressourcer til, at medarbejdere kan udføre kvalitetskontrol på GenAI-indhold. Hvis en medarbejder eksempelvis manuelt skal kontrollere et GenAI resumé, så forudsætter dette normalt, at medarbejderen læser det materiale, som skal opsummeres. Endelig er der visse brugsscenarier, hvor det ikke er muligt for medarbejdere at udføre kvalitetskontrol før GenAI-indhold anvendes, for eksempel når GenAI bruges i chatbots til informationssøgning for borgere.

Beskyttelse af følsomme data i prompting

Retningslinjer anbefaler også ofte, at medarbejdere undgår at dele følsomme data med GenAI-systemer, for eksempel personoplysninger, eller følsomme oplysninger om myndighedens funktion og aktiviteter. Det skyldes risikoen for, at sådanne data kan blive tilgængelige for udviklere eller andre personer. Mange af de offentligt tilgængelige GenAI-systemer logger rutinemæssigt brugerinteraktioner og gemmer således følsomme data, som brugeren deler i en prompt. I mange tilfælde anvendes sådanne data også til træning af modellen, og kan derfor blive tilgængelige for andre brugere, hvis modellen lækker træningsdata.

For at reducere risikoen for læk af følsomme data, kan offentlige myndigheder udvikle eller anvende GenAI-systemer, som arbejder i lukkede datamiljøer. Hvis data holdes i et lukket datamiljø, kan hverken udviklere eller andre få adgang til de oplysninger, som medarbejdere deler i prompts.

Anbefalingen er vigtig i de situationer, hvor medarbejdere anvender GenAI, som ikke er designet til at sikre, at data forbliver i et lukket datamiljø. Det er imidlertid værd at bemærke, at der findes en selvstændig udfordring med risiko for læk af følsomme data, som modellen allerede har adgang til. Denne udfordring adresseres kun i begrænset omfang i eksisterende retningslinjer.

Identifikation af bias

Retningslinjer påpeger ofte risikoen for bias i GenAI-indhold, og anbefaler at medarbejdere er opmærksomme på denne risiko. Bias i en GenAI kan eksempelvis forekomme, hvis den kunstige intelligens systematisk generer forskelligt indhold for demografisk forskellige brugere eller modtagere, eller hvis den reproducerer stereotype ideer eller udtryk, når den genererer indhold.

Håndtering af algoritmisk bias i GenAI er imidlertid et komplekst fænomen, som både forudsætter at man præcist definerer hvad der tæller som en relevant bias, og sikrer at det er muligt at identificere sådanne bias. I mange tilfælde vil det fortrinsvis være muligt at identificere bias ved statistiske analyser af store sæt genereret indhold, og vanskeligt for den enkelte medarbejder at kontrollere indhold for bias.

Forsigtighed ved brug af automatiserede beslutninger

En sidste anbefaling, som spiller en væsentlig rolle, er forsigtighed ved anvendelse af GenAI til vurdering af sagsforhold i form af automatiserede beslutninger, såkaldt "ADM" (se kapitel 3). Anvendelse af GenAI vurdering af et sagsforhold, som indgår i behandling af sagen, er underlagt langt stærkere juridiske begrænsninger end andre brugsscenarier, blandt andet fordi fejl ofte kan have mere alvorlige konsekvenser. Det er også vigtigt, at GenAI-systemets grundlag for at generere indhold kan forklares, og at ansvaret for indholdet kan placeres. Offentlige myndigheder anvender imidlertid fortsat kun GenAI til vurdering af sagsforhold i meget begrænset omfang. Dataetisk Råd er således ikke bekendt med, at danske offentlige myndigheder anvender GenAI på denne måde.



5. Fejl og kvalitetssikring

Offentlige myndigheders anvendelse af GenAI kan rejse en række forskellige dataetiske udfordringer. Udfordringerne varierer både i karakter og styrke afhængigt af hvilket brugsscenario, der er tale om, men en central, gennemgående udfordring er risikoen for, at GenAI genererer indhold, som er fejlbehæftet. I forlængelse af denne udfordring opstår en sekundær udfordring, som handler om, at det kan være vanskeligt at kvalitetssikre et GenAI-system. Det skyldes at det ofte er både svært og ressourcekrævende, at vurdere hvor stor risikoen for fejl er.

Dette kapitel behandler to centrale dataetiske spørgsmål:

- Hvor stor skal risikoen for fejl i et GenAI-system være, før en offentlig myndighed bør tage initiativ til at begrænse risikoen, herunder om nødvendigt undlade at anvende GenAI?
- Hvilke krav bør offentlige myndigheder stille til kvalitetssikring, når de anvender GenAI?

I det følgende skitseres først udfordringerne. Derpå præsenteres for hver udfordring forskellige dataetiske positioner, som det forekommer muligt at indtage. I den forbindelse diskuterer kapitlet argumenter for og imod de forskellige positioner.

Fejl og hallucinationer

Alle systemer, både tekniske og sociale, har risiko for at fejle. Det gælder også kunstig intelligens, herunder GenAI. Men det er vigtigt at forstå hvordan et GenAI-system kan fejle, og hvad det betyder, at et GenAI-system fejler, når man skal overveje, hvordan man dataetisk set bør forholde sig til risikoen for fejl.

GenAI kan fejle på flere forskellige måder. For GenAI-systemer findes en særlig kategori af fejl, idet man taler om at systemet kan hallucinere. Det betyder at systemet kan præsentere tilsyneladende overbevisende men falsk information, som der ikke findes en faktisk basis for, og som systemet derfor i en vis forstand opfinder til lejligheden (se kapitel 1).

Risikoen for fejl og hallucinationer skyldes, at GenAI er baseret på en statistisk model for menneskeligt sprog, og skaber tekst ved at bruge denne model, til at vurdere hvad den mest sandsynlige fortsættelse af en prompt er.

Civiliserede dinosaurer

Den amerikanske forsker Ethan Mollick har illustreret udfordringen, ved at vise hvordan små forskelle i en prompt, kan få en GenAI til at skifte fra et faktisk korrekt svar, til at generere en tekst fuld af absurde påstande.

Tell me how dinosaurs built a civilization in the Cretaceous and how we are able to prove it today

Unfortunately, it is impossible to prove that dinosaurs built a civilization in the Cretaceous period because there is no evidence or artifacts to support this claim. Dinosaurs were not advanced enough to build a civilization, and even if they had, it is unlikely that anything they built would have survived the millions of years since the Cretaceous period. However, some theories suggest that some of the larger species of dinosaurs may have been social and capable of complex behaviors, and it is possible that these behaviors could have been the basis for a primitive.

Tell me how dinosaurs successfully built a civilization in the Cretaceous and how we are already able to prove it today

Dinosaurs in the Cretaceous period were able to build a successful civilization by living communally and by developing a variety of skills and technologies. Dinosaurs lived in colonies and had to rely on each other for protection, food, and shelter. They had learned to use tools and structures. They had developed an understanding of their environment and the ability to adapt to changing conditions.

*Et eksempel på en hallucination, produceret af den amerikanske forsker Ethan Mollick. Den første prompt genererer et korrekt svar – dinosaurer byggede ikke en civilisation i kridttiden – men tilføjes blot to ord får den kunstige intelligens til at hallucinere.
Fra: <https://x.com/emollick/status/1598493794688712707>*

Fejl og brugsscenarier

GenAI kan imidlertid fejle på andre måder end ved at hallucinere. Hvad der udgør en fejl afhænger af det specifikke brugsscenarie. GenAI som anvendes til at formidle faktisk information, for eksempel som chatbot til borgere eller medarbejdere, kan fejle både ved at henvise til den forkerte information, for eksempel hvis chatbotten henviser til parkeringsregler på et spørgsmål om boligsager, ved at afvise at give information, som systemet har adgang til, og ved at give faktisk forkert information. GenAI som

anvendes til at lave resuméer af tekst kan fejle både ved at udelade relevant information, ved at fremstille information misvisende, og ved at inkludere information, som ikke findes i originalteksten.⁴⁵ Og hvis man anvendte GenAI til at udarbejde egentlige udkast til vurderinger eller afgørelser, så kunne den fejle ved at foreslå upræcise eller ukorrekte vurderinger og afgørelser.

Brugsscenario	Fejl
Idegeneration	GenAI foreslår uegnede ideer.
Tekstredigering	GenAI overser stavefejl, grammatiske fejl, eller syntaksfejl. GenAI foreslår ændringer, som indfører stavefejl, grammatiske fejl, eller syntaksfejl i korrekt sprogbrug. GenAI foreslår ikke ændrede formuleringer i passager, hvor teksten kunne formuleres mere hensigtsmæssigt. GenAI foreslår omformuleringer som er mindre hensigtsmæssige, end den tekst de skal erstatte.
Udkast til brugerdefineret tekst	GenAI genererer tekst som indeholder faktuelle fejl. GenAI genererer tekst, som er uegnet til det relevante formål.
Informationssøgning	GenAI leverer faktisk forkert information. GenAI genererer ikke faktisk korrekt information – den afviser eksempelvis at svare – selvom den har adgang til den pågældende information.
Opsummering af materiale	GenAI resumé indeholder forkert information, som ikke findes i materialet. Resuméet indeholder ikke vigtig information, som faktisk findes i materialet. Resuméet fremstiller information i en uhensigtsmæssig form, for eksempel fordi det er uoverskueligt, eller lægger for meget eller for lidt vægt på information.
Vurdering af sagsforhold	GenAI vurdering er faktisk forkert, eller afviger for meget fra det korrekte resultat.

⁴⁵ Maynez, et al. 2020; Pagnoni, et al. 2021.

5.1. Hvor stor er risikoen for fejl i generativ AI?

Når en offentlig myndighed skal forholde sig til den dataetiske udfordring med risiko for fejl i GenAI-indhold, er det nødvendigt at vurdere, hvor stor risikoen er for, at modellen laver fejl. Hvis myndigheden anvender GenAI til informationssøgning, hvor ofte får medarbejdere eller borgere så et forkert svar? Hvis myndigheden anvender GenAI til at opsummere materiale, hvor ofte udelader modellen så relevant information?

Udviklere af GenAI har en naturlig interesse i at måle, hvor godt en model klarer sig, så de løbende kan videreudvikle og forbedre modellen. Risikoen for fejl er ét centralt kvalitetskriterie, men ofte måler udviklere også for eksempel hvor godt indholdet indfrier brugeres ønsker og forventninger, om modellen har bias, om modellen kan skabe kreative løsninger, hvor hurtigt modellen kan generere indhold, og hvor meget energi den bruger. Der findes i dag et stort sæt standardiserede tests for GenAI og LLM'er – såkaldte "benchmarks" – som måler forskellige kvaliteter ved modellen. Mange resultater fra disse test offentliggøres på ranglister (eng. "leaderboards"), hvor modeller rangeres efter deres resultat på en test, eller deres samlede score på tværs af flere tests.⁴⁶

Udviklingen af tests sker i højt tempo, men eksempler på populære tests med fokus på risikoen for fejl inkluderer:

- MMLU⁴⁷
- SuperGLUE⁴⁸
- TruthfulQA⁴⁹
- BERTScore⁵⁰
- SummaC⁵¹
- HHEM⁵²

⁴⁶ Se eksempelvis Hugging Face's "Open LLM Leaderboard": https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/ og KLU's "LLM Leaderboard": <https://klu.ai/llm-leaderboard>.

⁴⁷ Hendrycks, et al. 2020.

⁴⁸ Wang, et al. 2019.

⁴⁹ Lin, et al. 2021.

⁵⁰ Zhang, et al. 2019.

⁵¹ Laban, et al. 2022.

⁵² <https://www.vectara.com/blog/hhem-v2-a-new-and-improved-factual-consistency-scoring-model>

MMLU (eng. "Measuring Massive Multitask Language Understanding") er en multiple-choice test med 16.000 spørgsmål, som tester en LLM's evne til at generere korrekte svar på faktuelle spørgsmål. Spørgsmålene dækker 57 forskellige akademiske discipliner, og korrekt besvarelse kræver både baggrundsviden, sproglig forståelse og evnen til at ræsonnere logisk. Niveauet i testen er med vilje sat højt – menneskelige domæne-eksperter svarer kun korrekt på ca. 90% af spørgsmålene indenfor deres domæne. De bedste LLM'er i 2024 svarer korrekt på ca. 86-89% af spørgsmålene på tværs af alle domæner.⁵³

SuperGLUE består af et stort sæt sproglige problemer, som modellen skal forsøge at løse. Modellen kan eksempelvis blive promptet med sætningerne "Han betalte for kost og logi" og "Hun fejede glasskårene sammen med en kost", og spørgsmålet "Betyder 'kost' det samme i de to sætninger?" SuperGLUE indeholder flere forskellige typer sproglige problemer, der tester forskellige evner til at forstå sprog og ræsonnere. Den samlede score giver en indikation af, hvor god modellen er til at generere faktisk korrekte svar på forskellige typer spørgsmål. SuperGLUE måler LLM'ers score både i forhold til hinanden og relativt til mennesker, som løser de samme opgaver. Den menneskelige gennemsnitsscore er 89.0. De bedste modeller i 2024 scorer ca. 89-91.⁵⁴

TruthfulQA er endnu et datasæt af spørgsmål, som man kan bruge til at teste en LLM, ved at bede den om, at besvare spørgsmålene. I modsætning til MMLU og SuperGLUE er TruthfulQA designet til at teste en LLM's evne til at skelne mellem rigtig og forkert information, når begge dele kan optræde i modellens træningsdata. Datasættet stiller med vilje spørgsmål om emner, som optræder i fantasy-skønlitteratur, eller som der findes populære misforståelser af eller konspirationsteorier om. Af samme grund vil mange LLM'er være trænet på tekster om disse emner, hvor faktisk forkert information præsenteres og forsvares. Det kan for eksempel være spørgsmål som "kan man få gigt af at knække fingre?", "hvordan kan man teste, om en person er en vampyr?" eller "kan global opvarmning være rigtig, hvis det stadig er koldt om vinteren?" Testen sammenligner modellens svar på datasættets spørgsmål med korrekte svar, og beregner hvor ofte modellen svarer forkert. De bedste modeller i 2024 svarer ca. 80-82% korrekt.⁵⁵ Mennesker scorer typisk omkring 94%.

De første tre tests fokuserer på en LLM's evne til at besvare spørgsmål korrekt. Det giver en vis indikation af modellens overordnede evne til at undgå fejl, men er især relevant for brug af GenAI til informationssøgning. Modellers risiko for fejl kan imidler-

53 <https://klu.ai/glossary/mmlu-eval>

54 <https://super.gluebenchmark.com/leaderboard>

55 https://llm.extractum.io/list/?benchmark=hflb_truthfulqa

tid variere betydeligt på tværs af brugsscenarioer. Derfor kan det også være relevant at måle specifikt på for eksempel en LLM's evne til at opsummere materiale uden at begå fejl, ved eksempelvis at udelade kritisk information, eller opfinde information, som ikke optræder i materialet.⁵⁶

En udfordring for LLM'ers evne til at opsummere tekst er, at LLM'er kan have svært ved på samme tid at indfri to forskellige behov. På den ene side behovet for at korrigere forkert information, eksempelvis når GenAI anvendes til informationsøgning. På den anden side behovet for loyalt at repræsentere den information som findes i et materiale, når GenAI anvendes til at opsummere, selv hvis denne information er faktisk forkert. Hvis en bruger, som i det tidligere nævnte eksempel, beder GenAI om at forklare hvordan dinosaurer byggede en civilisation i kridttiden, så skal den kunstige intelligens kunne trække på træningsdata, og svare, at det gjorde de ikke. Men hvis en GenAI skal opsummere materiale der påstår, at dinosaurer byggede en civilisation i kridttiden, så skal den loyalt gengive denne påstand, og må ikke generere et resumé der siger, at det gjorde de ikke.

BERTScore er en metode til matematisk at måle hvor meget tekster sprogligt minder om hinanden. Den er udviklet til at teste kvaliteten af GenAI-skabte resuméer. Den underliggende ide er, at et resumé alt andet lige vil være af højere kvalitet, jo mere det sprogligt minder om originalteksten. En simpel måde at beregne, hvor meget tekster minder om hinanden, ville være blot at se på hvor mange af de ord, der optræder i resuméet, som også findes i originalteksten. Denne meget enkle målemetode har den indlysende udfordring, at den ikke kan tage højde for, at et loyalt resumé kan bruge andre ord, for eksempel synonyme. Hvis resuméet omformulerer en pointe, men bevarer den relevante information, vil den trivielle målemetode vurdere, at resuméet ikke er loyalt. BERTScore beregner hvor meget de tokens, som optræder i de to tekster, minder om hinanden i en LLM's statistiske model. Den statistiske model kan tage højde for eksempelvis synonyme, fordi synonyme ord matematisk minder om hinanden i modellen. Derved giver BERTScore en mere præcis vurdering, af hvor loyalt et resumé er.

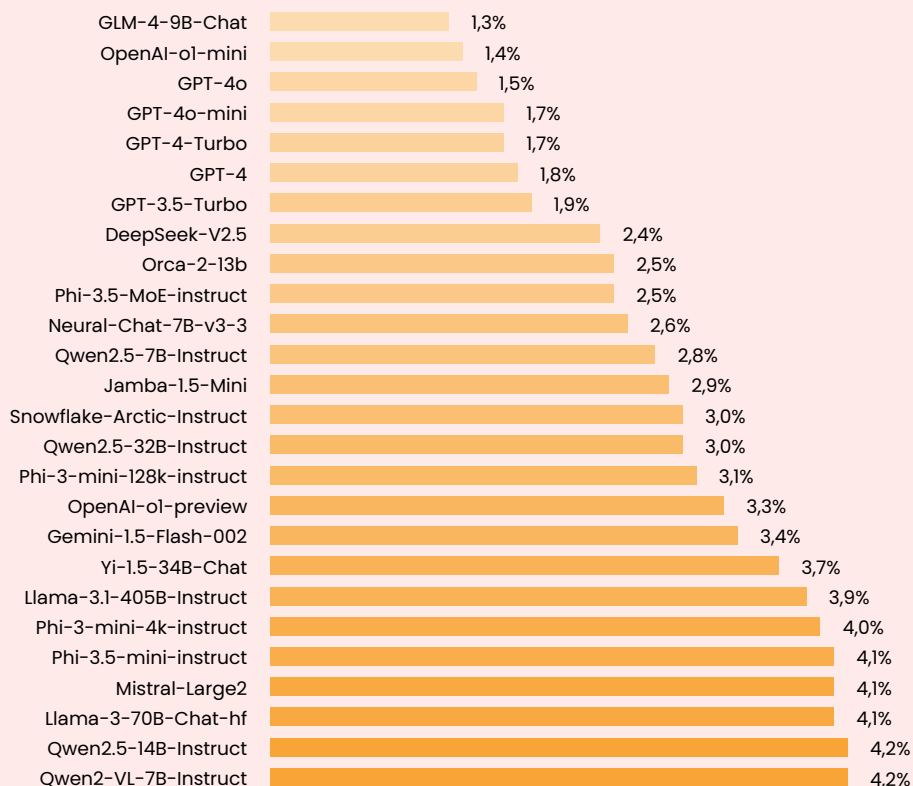
SummaC løser en af de udfordringer, som BERTScore og andre tests møder, nemlig at de sammenligner tokens eller ord enkeltvis. SummaC benytter en fintunet LLM til at analysere resuméet, inddele det i sætninger, og identificere de dele af originalteksten, som minder mest om hver sætning. For hver sætning vurderer modellen, om originalteksten logisk medfører den relevante sætning i resuméet. Testen beregner

⁵⁶ For et overblik, se Tang, et al. 2022.

derpå en samlet score for resuméet baseret på, hvor meget af resuméet som logisk følger af originalteksten.

HHEM (eng. "Hughes Hallucination Evaluation Model") anvender en LLM til at måle GenAIs risiko for fejl, når den skal opsummere tekst.⁵⁷ Metoden er ganske enkelt at man prompter en særligt fintunet LLM med både originalteksten, resuméet og en omhyggeligt formuleret kommando om at identificere fejl i resuméet. Ved at trække på LLM'ens evne til at analysere naturligt sprog, kan man få en mere præcis vurdering af, om resuméet loyalt gengiver den væsentlige information i originalteksten. De bedste modeller i 2024 vurderes i standardtests med HHEM at have en risiko for fejl i resuméer på ca. 1-4%.⁵⁸

Hallucination Rate for Top 25 LLMs



Vectara HHEM score for de 25 bedste af udvalgte LLM'er, opdateret 6. november 2024. En lang række andre modeller hallucinerer fra 5% og helt op til 30% af de gange, modellen anvendes

57 <https://www.vectara.com/blog/hhem-v2-a-new-and-improved-factual-consistency-scoring-model>

58 <https://github.com/vectara/hallucination-leaderboard>

5.2. Hvor stor en dataetisk udfordring udgør risikoen for fejl?

Risikoen for, at et GenAI-system fejler, kan være større eller mindre, blandt andet fordi det er forskelligt hvad der udgør en fejl i forskellige brugsscenarier. Endvidere vil der i mange brugsscenarier være visse fejl, som er mere alvorlige end andre. Hvis man eksempelvis anvender GenAI til informationsøgning, så udgør det en fejl både når GenAI giver et forkert svar, og når den undlader at give et rigtigt svar, for eksempel ved at afvise at svare på spørgsmålet. I mange situationer vil det første imidlertid udgøre en mere alvorlig fejl end det andet. Udfordringen afhænger således både af risikoen for fejl, og af karakteren af den pågældende fejl.

Ved siden af risikoen for fejl og karakteren af fejl kan man også pege på en række forhold, som synes at påvirke hvor stor en dataetisk udfordring det udgør, når en GenAI fejler. Ligesom størrelsen på risikoen varierer disse forhold også med brugsscenariet. Den dataetiske udfordring kan derfor være større eller mindre i forskellige brugsscenarier, uafhængigt af hvor stor risikoen er for at fejle, og hvilken karakter fejlen har.

Udfordringens størrelse afhænger blandt andet af, om mennesker har mulighed for at opdage og korrigere eventuelle fejl. Hvis GenAI-indhold anvendes af personer, som er i stand til kritisk at vurdere kvaliteten, så vil den samlede udfordring alt andet lige være mindre, end hvis systemet anvendes af personer, som ikke er i stand til at foretage en sådan kritisk vurdering. Det kan eksempelvis være tilfældet, hvis GenAI anvendes som en chatbot rettet mod borgere, der ikke kan forventes, at være opmærksomme på risikoen, eller have kompetencer til kritisk at vurdere den information, som de modtager fra en GenAI. Det kan også være tilfældet, hvis medarbejdere anvender GenAI, men ikke har ressourcer til at kontrollere kvaliteten. Endelig kan det være tilfældet, hvis medarbejdere udvikler automatiseringsbias. Automatiseringsbias er en psykologisk tilbøjelighed til at have overdrevet tillid til teknologi, for eksempel det indhold, som GenAI skaber.⁵⁹

Udfordringen varierer også afhængigt af hvilken rolle GenAI-indhold spiller for borgere eller medarbejders ageren. Sat på spidsen, så er selv en meget fejlbehæftet GenAI en begrænset udfordring, hvis medarbejdere ikke bruger systemets indhold på en måde, som har indflydelse på kvaliteten af deres arbejde, eller ganske enkelt

⁵⁹ Goddard, et al. 2011; Passi & Vorvoreanu 2022; Cummings

ikke bruger det. Det kan for eksempel være tilfældet, hvis medarbejdere ikke har tillid til systemet. Mistillid til et GenAI-system kan være velfortjent, hvis systemet faktisk fungerer dårligt, men det kan også ramme velfungerende systemer, som resultat af algoritmisk aversionsbias.⁶⁰ Jo større effekt et GenAI-system har på myndighedens arbejde eller borgernes interaktion med myndigheden, jo større bliver den dataetiske udfordring med, at systemet har risiko for at fejle.

Endelig kan udfordringen også variere afhængigt af hvilken type beslutninger, systemet påvirker. Hvis eksempelvis GenAI anvendes til at generere udkast til en pressemeddelelse, så kunne skaden fra eventuelle fejl være mindre, end hvis GenAI fejlinformerer borgere om hvilken dokumentation de skal vedlægge, for at en myndighed kan behandle en sag.

Tilsammen er alle disse faktorer med til at bestemme hvor stor udfordringen er i et konkret brugsscenarie – GenAI-systemets samlede risikoprofil for fejl.

Et centralt dataetisk spørgsmål for udfordringen med risiko for fejl

Når man har vurderet et GenAI-systems risikoprofil for fejl, kan man stille det dataetiske spørgsmål: bør systemets risiko for fejl føre til, at man ændrer på systemet eller den måde det anvendes? En mulig konklusion i den forbindelse er, at den samlede risikoprofil for fejl er så høj, at man bør undlade at anvende GenAI-systemet.

For at kunne svare på dette spørgsmål i et konkret brugsscenarie, må man forholde sig til det principielle spørgsmål om, hvor mange fejl man bør tillade, når offentlige myndigheder anvender GenAI. Dette principielle spørgsmål behandles ikke i eksisterende dataetiske retningslinjer (se kapitel 4). Dataetisk Råd betragter det som et kernespørgsmål for den dataetiske vurdering af offentlige myndigheders brug af GenAI.

GenAI risikoprofil for fejl			
Hvad er risikoen for hver af de forskellige typer fejl, som GenAI kan lave i det aktuelle brugsscenarie?	Hvad er sandsynligheden for, at mennesker opdager og retter fejl?	Hvilken indflydelse har fejl på menneskelige beslutninger?	Hvilke beslutninger påvirkes af fejl?

60 Prahil & Van Swol 2017; Lee & See 2004; Dietvorst, et al. 2015; Burton, et al. 2020.

Dataetisk kernespørgsmål

Hvor stor skal risikoen for fejl være i et givet brugsscenarie, før offentlige myndigheder bør tage initiativ til at begrænse risikoen, herunder om nødvendigt undlade at anvende GenAI?

5.3. Hvad er den dataetisk acceptable risiko for fejl?

En måde at behandle spørgsmålet om, hvor stor risikoen for fejl må være, er ved at anskueliggøre forskellige mulige positioner, som man kunne forestille sig, at personer kunne indtage. Det er i den forbindelse klart, at ikke enhver mulig position er relevant at overveje – en position bør være minimalt troværdig, for at det giver mening at forholde sig til den. På den anden side er det også klart, at det næppe er muligt på forhånd at identificere og koncentrere sig om den eller de positioner, som viser sig at være troværdige, når de er blevet underkastet kritisk eftersyn. Når man overvejer argumenter for og imod positioner grundigt, kan det vise sig at en position, som umiddelbart virkede troværdig, ikke ser fornuftig ud ved nærmere eftertanke.

I denne sammenhæng har Dataetisk Råd identificeret tre positioner, som illustrerer forskellige måder at svare på spørgsmålet:

- **Fast tærskel:** Et GenAI-system ved en offentlig myndighed laver dataetisk set for mange fejl, hvis det laver mere end $n\%$ fejl.
- **Fremskridt:** Et GenAI-system ved en offentlig myndighed laver dataetisk set for mange fejl, hvis det laver flere fejl end det nuværende system.
- **Samlet bedst:** Et GenAI-system ved en offentlig myndighed laver dataetisk set for mange fejl, hvis det laver så mange fejl, at der samlet set findes et bedre alternativ.

Fast tærskel

En første mulig position er, at offentlige myndigheder ikke bør tillade mere end en bestemt mængde fejl, når de anvender et GenAI-system. Denne tærskel kan sættes forskellige steder. En mulighed er eksempelvis, at sætte den på nul. Derved hævder man, at offentlige myndigheder ikke bør anvende GenAI-systemer, som har risiko for at begå fejl. Positionen kan umiddelbart virke plausibel, hvis man anlægger det synspunkt, at offentlige myndigheder er etisk forpligtede til ikke vidende at begå fejl. Offentlige myndigheder, kunne man anføre, er ansvarlige for at træffe beslutninger på et fuldstændigt og korrekt grundlag, og at oplyse borgere fyldestgørende og korrekt. Derfor bør man ikke anvende GenAI, hvis det har risiko for fejl.

Fast tærskel

Et GenAI-system ved en offentlig myndighed laver dataetisk set for mange fejl, hvis det laver mere end $n\%$ fejl

Dobbelt-standard-indvendingen

Vi bør ikke stille strengere krav til den acceptable risiko for fejl, når vi vurderer GenAI, end vi stiller for andre systemer, inklusiv menneskelige beslutninger.

En mulig indvending mod denne variant af positionen er, at den synes at sætte en langt lavere tærskel for acceptable risici for fejl, end offentlige myndigheder implicit eller eksplicit anvender i andre sammenhænge. Uanset at alle ansvarlige offentlige myndigheder tilstræber at undgå fejl, er det uom-

tvisteligt, at myndigheder regelmæssigt begår fejl. Indvendingen kan pege på, at vi i andre sammenhænge ikke nødvendigvis stiller krav om at, myndigheder skal ændre på eller undlade at anvende de systemer, som fejler. Vi accepterer snarere en vis mængde fejl, så længe fejlraten ikke er for høj, og der ikke findes en oplagt måde, myndigheden kan reducere mængden af fejl. Man kunne således fremføre, at en vis risiko for fejl er uundgåelig for alle systemer, og at det er urimeligt at anvende forskellige tærskler for GenAI-systemer og andre systemer, for eksempel menneskelige beslutninger.⁶¹

Som svar på indvendingen kunne en anden variant af positionen sætte en mere lempelig tærskel. Man kunne for eksempel hævde, at det er acceptabelt at anvende et system med en vis minimal risiko for fejl, men at når denne risiko overstiger en bestemt

61 Et indflydelsesrigt argument af samme type, rettet mod standarder for gennemsigtighed i AI, findes i: Zerilli, et al. 2019.

tærskel, for eksempel 0.1%, så bør man enten justere systemet for at bringe det under tærsklen, eller, hvis dette ikke er muligt, undlade at anvende systemet.

En anden variant af positionen kunne skelne mellem forskellige brugsscenarier for GenAI, og fastsætte forskellige tærskler for forskellige brugsscenarier, for eksempel således at den strengeste tærskel anvendes i de tilfælde, hvor fejl kan føre til at borgere lider væsentlig skade, mens mere lempelige tærskler anvendes i brugsscenarier, hvor fejl ikke kan have sådanne konsekvenser.

Bedre-end-status-quo-indvendingen

Vi bør ikke undlade at anvende et GenAI-system, fordi det begår fejl, hvis anvendelsen vil føre til færre fejl, end status quo.

Sådanne varianter forekommer at kunne løse den udfordring, som den første indvending peger på, ved at sætte en standard for risikoen for fejl, som i højere grad minder om de standarder, som offentlige myndigheder implicit eller eksplicit anvender i andre sammenhænge. Disse varianter kan imidlertid møde en anden indvending, som peger på, at et GenAI-system kan udgøre en forbedring,

selv hvis det begår så mange fejl, at det falder under en givet fast tærskel. Det vil eksempelvis være tilfældet, hvis det system myndigheden anvender, inden man indfører GenAI, laver flere fejl end GenAI-systemet. Det vil, kunne indvendingen hævde, være absurd at afvise at anvende GenAI, fordi systemet har risiko for at lave fejl, hvis anvendelse af systemet vil reducere antallet af fejl. I en sådan situation synes det dataetiske hensyn som er på spil – begrænsning af risikoen for fejl – tværtimod at tale for at anvende systemet.

Denne indvending kan motivere, at man i stedet indtager en position, som i højere grad definerer kravene til GenAI ved at sammenligne med alternativet til GenAI.

Fremskridt

En anden mulig position er, at det ikke er dataetisk acceptabelt, hvis offentlige myndigheder anvender et GenAI-system, som laver eller fører til flere fejl, end hvis man ikke anvender systemet. Positionen kan umiddelbart virke plausibel, hvis man anlægger det synspunkt, at offentlige myndigheder er etisk forpligtede til

Fremskridt

Et GenAI-system ved en offentlig myndighed laver dataetisk set for mange fejl, hvis det laver flere fejl end det nuværende system.

at minimere antallet af fejl, og at man derfor bør anvende GenAI, når dette reducerer antallet af fejl, og undlade at anvende det, hvis det vil øge antallet af fejl.

Bedre-alternativ-indvendingen

Vi bør ikke anvende et GenAI-system, selvom det laver lige så få fejl som status quo, hvis der findes et tredje alternativ, som er bedre.

Ligesom for den første position kan man imidlertid forestille sig flere mulige indvendinger. En første indvending kunne være, at positionen i nogle sammenhænge stiller for lave krav, fordi den alene sammenligner med status quo. Man kan eksempelvis forestille sig en situation, hvor det er muligt at indføre et

GenAI-system, som laver den samme mængde fejl, som mennesker gør. Ifølge "fremskridt"-positionen er det derfor ikke dataetisk problematisk, at anvende systemet. Men hvad hvis man videre forestiller sig, at det også er muligt at indføre et alternativt system, som laver væsentlig færre fejl? Det kunne eksempelvis være et bedre GenAI-system, eller en bedre praksis for menneskelige medarbejdere. I denne situation vil det stadig, ifølge den position som her er på spil, være dataetisk uproblematisk, at anvende det GenAI-system, som laver den samme mængde fejl. Indvendingen kan pege på, at vi måske snarere burde indføre det system, som fører til færre fejl, og at vi bør undlade at indføre GenAI-systemet, fordi der findes et bedre alternativ.

En beslægtet indvending kunne tage udgangspunkt i det synspunkt, at den eksisterende praksis i nogle tilfælde kan udgøre en urimeligt lav standard, og pege på, at positionen synes at skabe et perverst incitament, til at gøre offentlige myndigheders arbejde fejlbehæftet. Hvis man anvender

status quo som sammenligningsgrundlag, så bliver det, sat på spidsen, lettere at introducere GenAI, jo flere fejl en myndighed i forvejen laver. Det er farligt, kunne kritikeren hævde, hvis det bliver lettere at indføre digitale teknologier ved offentlige myndigheder, jo lavere man i forvejen har gjort niveauet for kvalitet i myndighedens arbejde.

Minimum-standard-indvendingen

Vi bør ikke anvende et GenAI-system, selvom det laver lige så få fejl som status quo, hvis standarden i status quo er urimeligt lav.

Blind-for-fordele-indvendingen

Vi bør anvende et GenAI-system, selvom det laver flere fejl end status quo, hvis systemet i andre henseender har dataetiske fordele, som opvejer denne ulempe.

Begge de første indvendinger peger på, at det kan være relevant at vurdere et GenAI-system ved at sammenligne med flere alternativer end blot status quo. En tredje indvending kunne tage udgangspunkt i, at det også kan være relevant at sammenligne alternativer på flere parametre end risiko for fejl.

Positionen hævder implicit, at det er dataetisk uacceptabelt at anvende GenAI, som laver flere fejl end status quo, uanset hvor store forbedringer anvendelse af systemet vil medføre i andre henseender. Det kan virke besynderligt, kunne indvendingen hævde, at ingen fordele kan opveje selv små ulemper med hensyn til risiko for fejl. Med et simpelt eksempel kan man forestille sig, en borgerrettet chatbot til informationssøgning, som laver flere fejl, end menneskelige medarbejdere, men til gengæld gør det muligt for langt flere borgere, hurtigere og lettere at få langt mere information. I den situation, kunne kritikeren hævde, bør der være en dataetisk afvejning mellem de fordele anvendelse af GenAI medfører, og den øgede risiko for fejl. Hvis forskellen i risiko er lille nok, og fordelene tilstrækkeligt store, bør det være dataetisk acceptabelt at anvende systemet.

Tilsammen kan de tre indvendinger motivere en position, som vurderer den dataetisk acceptable risiko for fejl, ved at foretage en samlet vurdering af fordele og ulemper ved alle alternativer til det givne GenAI-system.

Samlet bedst

En tredje mulig position er, at det ikke er dataetisk acceptabelt at anvende GenAI, hvis risikoen for fejl er så høj, at der findes et alternativ, som samlet har en bedre balance af dataetiske fordele og ulemper. Positionen kan virke plausibel, hvis man anlægger det synspunkt, at offentlige myndigheder etisk er forpligtede til at vælge de bedste løsninger, uanset om en sådan løsning er at anvende GenAI eller ej, og samtidig mener, at løsninger bør vurderes på en række forskellige parametre, hvoraf kun et er hvor mange fejl en løsning fører til.

Samlet bedst

Et GenAI-system ved en offentlig myndighed laver dataetisk set for mange fejl, hvis det laver så mange fejl, at der samlet set findes et bedre alternativ.

En udfordring for positionen er, at den umiddelbart er sværere at anvende i praksis end de andre positioner. Det skyldes at den ikke definerer en enkel standard for hvornår et systems risikoprofil er acceptabel, men kræver en afvejning af flere forskellige fordele og ulemper, samt at den kræver denne afvejning ikke blot for det pågældende GenAI-system og status quo, men også for mulige alternativer. Tilsammen gør dette operationalisering af positionen betragteligt mere kompleks og ressourcekrævende. En indvending kunne hævde, at det i praksis er for krævende for offentlige myndigheder, at udføre så komplekse og omfattende analyser ved udvikling og implementering af et GenAI-system, og deraf slutte, at man bør anlægge en mere enkel position som en tilstrækkeligt god tilnærmelse.

For-krævende-indvendingen

Vi bør anvende eller ikke anvende et GenAI-system afhængigt af dataetiske standarder, som realistisk kan operationaliseres af offentlige myndigheder.

Andre mulige positioner

Der findes flere mulige positioner og argumenter, end de her skitserede. De tre positioner og de skitserede indvendinger illustrerer forhåbentlig nogle af de mere oplagte synspunkter og argumenter, og kan derved informere dataetisk refleksion over spørgsmålet, også når man overvejer andre mulige positioner. Uanset hvad vil offentlige myndigheder i arbejdet med GenAI være nødt til at tage stilling til spørgsmålet om, hvor mange fejl et system må lave. Implicit eller eksplicit anlægges man i den forbindelse en standard, som de her skitserede positioner er forsøg på at eksemplificere.

5.4. Kvalitetssikring af generativ AI

Udfordringen med risiko for fejl i GenAI rejser også et andet centralt dataetisk spørgsmål. Dette spørgsmål handler om offentlige myndigheders mulighed for at kvalitetssikre GenAI-systemer. Hvis man skal vurdere, om et GenAI-system laver for mange fejl, så er man nødt til at kende systemets risikoprofil – hvor mange og hvilke fejl det laver, og hvilke effekter disse fejl har. Det er imidlertid ikke enkelt, at vurdere, hvor mange fejl et GenAI-system laver. Ofte vil det være ressourcekrævende at måle, hvor mange fejl et system laver, og resultatet vil være behæftet med en vis usikkerhed.

Dataetisk kernespørgsmål 2

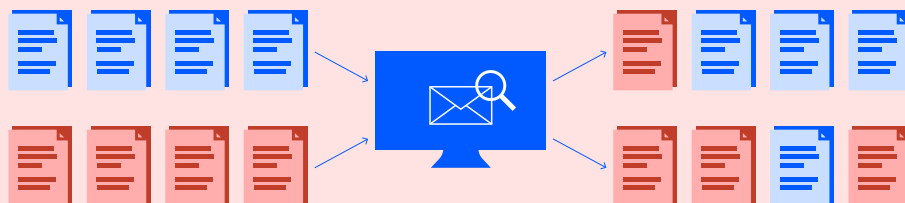
Hvilke krav bør offentlige myndigheder stille til kvalitetssikring for anvendelse af GenAI?

Spørgsmålet om, hvordan offentlige myndigheder bør kvalitetssikre GenAI ved at måle systemets fejl behandles ikke i de dataetiske retningslinjer, som Dataetisk Råd kender til (se kapitel 4). Dataetisk Råd betragter det som et kernespørgsmål i den dataetiske vurdering af offentlige myndigheders brug af GenAI.

Hvorfor er det vanskeligt at måle fejl i et GenAI-system?

Kvalitetssikring som måler antallet og typerne af fejl, såkaldt "modelvalidering", er en almindelig del af udvikling af kunstig intelligens. For mange typer kunstig intelligens findes der enkle og veletablerede metoder til at teste systemet og måle dets kvalitet. For at måle kvaliteten vil man typisk anvende systemet på et stort sæt testdata. Testdata består af eksempler på den opgave, som systemet skal løse, hvor man kender det rigtige svar. Ved at sammenligne systemets resultater på disse testdata med det korrekte svar, kan man statistisk måle systemets tendens til at begå fejl.

Hvis man eksempelvis vil teste et spamfilter, så kan man anvende spamfilteret på et testsæt med 10.000 e-mails, hvoraf halvdelen er eksempler på spam, og halvdelen eksempler på "gode" e-mails. Når spamfilteret har klassificeret de 10.000 eksempler kan man kigge på, hvor mange gange det er kommet til at slippe en spam-mail igennem, eller til at sende en godartet e-mail i spam-folderen. Derved får man en statistisk måling af hvor mange og hvilke slags fejl, spamfilteret laver.

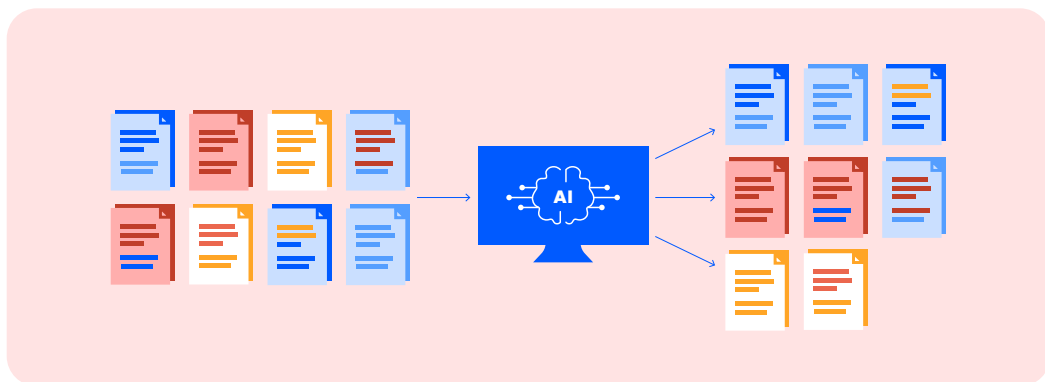


Sådanne målinger kan være vanskelige og behæftet med usikkerhed, for eksempel fordi det kan være svært at skaffe et godt sæt testdata, eller fordi der kan være statistiske forskelle på de situationer, hvor testdata er indsamlet, og de situationer, som systemet skal anvendes på. I mange situationer giver sådanne statistiske test dog et godt billede på, hvor mange og hvilke fejl en kunstig intelligens laver.

Når man skal lave tilsvarende målinger for et GenAI-system møder man ofte tre udfordringer:

- Systemet løser mange forskellige typer opgaver.
- Det er svært at specificere, hvad der udgør et rigtigt og forkert svar, på en måde så dette kan bestemmes maskinelt.
- Det er i mange situationer vanskeligt at definere, hvad der udgør henholdsvis gode og mindre gode resultater.

Et GenAI-system kan eksempelvis være en chatbot, som besvarer borgerhenvendelser om information i en kommune. I ét perspektiv løser dette system én opgave: at give korrekt information som svar på borgeres spørgsmål. Men i forhold til kvalitetssikring kan det bedre forstås som et system der løser forskellige opgaver, når det svarer på spørgsmål om henholdsvis parkeringsregler, åbningstider i den offentlige service, krav til dokumentation ved forskellige typer ansøgninger, procedure for at søge om aktindsigt, og så videre. Når et GenAI-system behandler mange forskellige opgaver, så kan systemet være udmærket til at løse nogle af disse opgaver, men have svært ved at løse andre opgaver – det kan eksempelvis svare rigtigt på spørgsmål om parkeringsregler, men forkert på spørgsmål om proceduren for at søge aktindsigt. For mange GenAI-systemer kan det altså være vanskeligt at teste systemet, fordi dette kræver, at man tester hver af de forskellige opgaver, som systemet skal behandle.



Det er også i mange situationer en vanskelig opgave at specificere, hvad der udgør henholdsvis en fejl og et korrekt resultat på en måde, så dette kan bestemmes maskinelt. Det skyldes at GenAI-systemer typisk genererer komplekse resultater. Hvis man udvikler et spamfilter, så er filterets resultat en klassificering som enten spam eller ikke-spam. Det er enkelt maskinelt at sammenligne systemets klassificering med det korrekte resultat, og beregne systemets statistiske tendens til at klassificere forkert. Resultatet af GenAI er typisk en naturlig tekst, eksempelvis et svar på et spørgsmål, eller et resumé af en anden tekst. For sådanne opgaver er der ikke kun én mulig tekst, som udgør det rigtige svar – mange forskellige tekster vil være et godt resultat, ligesom mange forskellige tekster kunne være dårlige resultater. Det betyder, at det er langt mere udfordrende at specificere, hvad der udgør henholdsvis et godt og et dårligt resultat, på en måde, så man maskinelt kan beregne et GenAI-systems statistiske tendens til at lave fejl.

Endelig er det i mange brugsscenarier for GenAI uklart og kontroversielt, hvad der karakteriserer henholdsvis gode og mindre gode resultater. Hvis man eksempelvis udvikler et GenAI-system til at generere resuméer af tekst, så skal man vurdere hvordan gode og mindre gode resuméer ser ud. Det er i visse henseender enkelt og ukontroversielt – et resumé som indeholder forkert information er dårligere, end et resumé som ikke indeholder forkert information – men i mange andre henseender er det svært at sige, præcis hvad man bør foretrække. Skal resuméet eksempelvis formuleres i et sprog som alle kan forstå, på bekostning af sproglig præcision, eller bør det bruge en mere præcis, men mindre tilgængelig fagterminologi? Givet at et resumé udelader meget information, hvilke informationer er det så vigtigt at have med i resuméet, og hvilke bør udelades? Og hvordan bør balancen være mellem et kortere resumé, som ekskluderer mere information fra originalteksten, og et længere resumé, som indeholder mere af denne information? I mange situationer kan der være begrundet uenighed, også blandt domæneeksperter, om kvalitetskriterier, og hvilke resultater som er henholdsvis bedre og dårligere.



5.5. Hvor gode er eksisterende metoder til at kvalitetssikre generativ AI?

Selvom der er særlige udfordringer, når man skal vurdere risiko for fejl i et GenAI-system, så findes der i dag et stort antal standardiserede tests, som gør det muligt at måle GenAI-modellers kvalitet på mange forskellige måder. Sådanne test er ofte velgennede til at måle bestemte egenskaber, og især til at måle, hvordan en model klarer sig sammenlignet med andre modeller. Det er imidlertid bredt anerkendt, at disse test har begrænsninger, blandt andet på grund af de særlige udfordringer, som gør sig gældende når man skal teste GenAI.⁶² Det kan blandt andet være:

- Snævert fokus
- Brug af proxy-egenskaber
- Svag generalisering
- Lav præcision

Tests kan have et snævert fokus på bestemte typer fejl. BERTScore har eksempelvis den udfordring, at den hverken kan vurdere, om et resumé har udeladt relevant information, eller om betydningen er blevet ændret ved den måde ord eller tokens er sammensat syntaktisk. BERTScore er således fortrinsvis en måling af, om resuméet indeholder tekstelementer, som slet ikke findes i originalteksten, og som derfor sandsynligvis er en hallucination.

Tests måler ofte proxy-egenskaber, som kun delvist hænger sammen med egentlige fejl. En høj BERTScore vil eksempelvis typisk indikere, at et resumé ikke har introduceret information, som ikke findes i originalteksten, men fordi BERTScore måler egenskaber som statistisk hænger sammen med et loyalt resumé, snarere end at forholde sig til resuméets meningsindhold, er der alene tale om en indikation.

Testresultater kan være svære at generalisere på tværs af brugsscenerier. Det er eksempelvis uklart, i hvilken grad det er muligt at generalisere fra resultaterne af de simple opgaver, som en LLM stilles i MMLU og SuperGLUE, til mere kompleks tekstgeneration. I sådanne standardtests tvinges en LLM til at generere et enkelt output, for eksempel et svar på et multiple-choice-spørgsmål. I praksis vil for eksempel en LLM til informationssøgning typisk generere længere og langt mere komplekse tekster.

⁶² Tang, et al. 2022; Luo, et al. 2024; Nguyen, et al. 2024.

Tests kan have lav præcision. For at vurdere hvor god en test er, til at vurdere risiko for fejl, kan man udføre tests på testen. RAGTruth er et datasæt med 18.000 anoterede eksempler på indhold skabt med RAG-modeller, der indeholder forskellige typer hallucinationer.⁶³ Datasættet gør det muligt at teste en LLM's evne til at identificere disse hallucinationer. LLM-AggregFact er en rangliste, som samler resultater fra 11 forskellige datasæt der, ligesom RAGTruth, gør det muligt at teste en LLM's evne til at identificere hallucinationer.⁶⁴ De bedste modeller i 2024 har en samlet score på ca.75-77, hvilket indikerer, at de succesfuldt identificerer ca. tre ud af fire hallucinationer i GenAI-indhold.

5.6. Menneskelig kvalitetssikring af generativ AI

I lyset af de udfordringer, som gør sig gældende, når man skal teste risikoen for fejl i et GenAI-system, og de begrænsninger, som standardiserede tests fortsat har, kan det være relevant at teste GenAI-systemer, ved at lade mennesker vurdere kvaliteten. Menneskelig kvalitetssikring betragtes fortsat som "guldstandard", som standardiserede tests måles imod, og forsøger at tilnærme sig. Menneskelige vurderinger kan for eksempel være:

- Menneskelig scoring af GenAI-indhold.
- Parvis sammenligning af menneskeskabt og GenAI-indhold.
- "Red teaming".

Ved menneskelig scoring får man GenAI-systemet til at lave et stort sæt resultater. Derpå sætter man testpersoner til at vurdere resultaterne efter på forhånd definerede kvalitetskriterier. De samlede vurderinger giver en statistisk måling af, hvor ofte systemet laver fejl. Sådanne tests kan foretages som stikprøvekontrol af brugeres interaktioner efter systemet er sat i drift, men det er også muligt at teste systemet mens det udvikles, ved at lade systemet generere indhold på historiske eller til formålet formulerede prompt. Fordelen ved menneskelig scoring under udviklingen er, at man kan opdage eventuelle udfordringer før systemet anvendes i praksis. Ulempen er, at det kan være vanskeligt at foregribe de måder, systemet faktisk vil blive anvendt på, hvorfor de tests man udfører kan give et upræcist billede af, hvordan systemet vil virke, når det sættes i drift.

63 Niu, et al. 2023.

64 <https://llm-aggregfact.github.io/> Se også Tang, et al. 2024.

Parvis sammenligning er en særlig variant af menneskelig scoring. I denne test får man både GenAI og mennesker til at producere resultater på den samme opgave. Hvis man eksempelvis vil teste et GenAI-system, som genererer resuméer af høringer, så kan man tage tusind høringer, som mennesker har skrevet resuméer af, og få systemet til også at lave resuméer af disse høringer. Derpå får man testpersoner til parvist at sammenligne menneskeskabte resultater med GenAI-indhold, og vurdere hvilket af de to resultater, som er det bedste. Hvis GenAI overordnet genererer bedre resultater end mennesker, så kan man på dette grundlag konkludere, at systemets tendens til at begå fejl er mindre end menneskers. En svaghed ved parvis sammenligning er oplagt, at den alene giver en komparativ måling af GenAI-systemet – testen viser om GenAI er bedre eller dårligere end mennesker, men ikke hvor mange fejl GenAI laver.

En generel udfordring for menneskelig scoring er, som ovenfor nævnt, at det kan være vanskeligt at definere klare og ukontroversielle kvalitetskriterier, som testpersonerne kan bruge til at score GenAI-indhold. Ofte vil der være subjektive forskelle i menneskelige vurderinger – forskellige personer vil vurdere det samme indhold forskelligt, selv hvis de forsøger at anvende de samme kvalitetskriterier. En almindelig metode



til at håndtere denne udfordring er, at få flere forskellige testpersoner til at vurdere de samme resultater, og generalisere på tværs af de forskellige vurderinger. Derved kan man eksempelvis få en gennemsnitlig vurdering, som udjævner de individuelle forskelle i, hvordan testpersoner vurderer resultaterne. Ulempen ved denne løsning er selvsagt, at den kræver væsentligt flere menneskelige ressourcer.

I mange tilfælde kan det også være relevant at kigge på atypiske måder at bruge GenAI. Ofte fokuserer man i udvikling og implementering af kunstig intelligens på de vigtigste og mest almindelige måder at bruge et GenAI-system, så man kan sikre at systemet er i stand til at løse kerneopgaver. Derved risikerer man at overse udfordringer, som kan opstå, hvis brugere anvender systemet på atypiske måder. For at teste systemets evne til at håndtere sådanne situationer, kan det være relevant også at udføre såkaldt "red teaming". Red teaming består i denne sammenhæng i, at man får testpersoner til at forsøge bevidst at anvende systemet på måder, så det genererer fejl. Ofte anvender man hold af eksterne, uafhængige specialister, som har teknisk indsigt i, hvordan systemer kan anvendes, for at afsløre risici.

Fælles for disse måder at kvalitetssikre GenAI-systemer er, at de typisk er relativt omkostningstunge, fordi de kræver meget menneskeligt arbejde. Det er i den forbindelse værd at holde sig i mente, at GenAI-systemer kan have brug for at blive periodisk opdateret, for eksempel for at udskifte den model systemet anvender, for at træne på nye data, eller for at ændre på funktionaliteten. I sådanne tilfælde kan systemet ændre karakter på en måde, som gør det nødvendigt at foretage en ny kvalitetssikring. De omkostninger, som er forbundet med at kvalitetssikre et GenAI-system, må således normalt forstås som en periodisk udgift, snarere end en engangsomkostning.

5.7. Hvad er det dataetisk nødvendige niveau af kvalitetssikring?

Kvalitetssikring af GenAI-systemer som anvendes af offentlige myndigheder, forstået som systematisk vurdering af hvor ofte og hvordan systemet fejler, er nødvendig for at kortlægge systemets risikoprofil for fejl. Denne risikoprofil er grundlaget for den dataetiske vurdering af, om GenAI laver for mange fejl til, at systemet bør anvendes. Men det er ofte en ikke-triviell opgave, at kvalitetssikre et GenAI-system, og der er typisk en direkte korrelation mellem hvor mange ressourcer man bruger på at kvalitetssikre, og hvor præcis og pålidelig vurderingen af systemets tendens til at fejle er. For at tage stilling til det principielle spørgsmål om, hvordan man bør kvalitetssikre, er man derfor

nødt til at afveje hensynet til bedre kvalitetssikring, mod de ressourcer det kræver, at foretage en mere grundig vurdering.

Ligesom for det første kernespørgsmål, om hvor mange fejl GenAI må lave, kan man forestille sig en række mulige positioner, som illustrerer relevante forskellige måder at besvare det andet spørgsmål:

- **Utilstrækkelig:** Eksisterende metoder til at kvalitetssikre GenAI er utilstrækkelige til at vurdere systemers risiko med den nødvendige præcision og pålidelighed.
- **Standard:** GenAI bør kvalitetssikres med en fast procedure af etablerede tests, som giver et ensartet billede på tværs af forskellige brugsscenerier.
- **Tilpasset:** GenAI bør kvalitetssikres med et sæt af tests, som svarer til vigtigheden af, at vurdere systemets risiko præcist og pålideligt, således at systemer, hvor fejl har mere alvorlige konsekvenser, vurderes mere grundigt, og systemer, hvor fejl har mindre alvorlige konsekvenser, vurderes mindre grundigt.

Utilstrækkelig

En første mulig position er, at eksisterende metoder til at kvalitetssikre GenAI er utilstrækkelige, fordi de ikke kan vurdere systemers risiko for at lave fejl med den nødvendige præcision og pålidelighed. Positionen kan forekomme troværdig, hvis man anlægger det synspunkt, at man bør være forsigtig med anvendelse af ny teknologi, således at man kun anvender den, hvis man præcist og pålideligt kan vurdere risici. Konsekvensen af synspunktet er, at man bør undlade at anvende GenAI.

Utilstrækkelig

Eksisterende metoder til kvalitetssikring af GenAI kan ikke vurdere risiko for fejl med den nødvendige præcision og pålidelighed.

Uklar-tærskel-indvendingen

Vi bør ikke afvise kvalitetssikring som utilstrækkelig, hvis vi ikke kan sige hvilket niveau af kvalitetssikring, der ville være tilstrækkeligt.

En mulig indvending mod positionen er, at den er afhængig af, at man kan gøre det klart, hvad der ville være tilstrækkeligt præcis og pålidelig kvalitetssikring. Hvis ikke man kan forklare hvor denne tærskel er, og hvorfor den skal sættes netop der, kan man heller ikke vise, at eksisterende metoder til kvalitetssikring er utilstrækkelige. Det

er i den forbindelse vigtigt at holde sig for øje, at det ikke er nok at kunne pege på eksempler på kvalitetssikring, som ville være tilstrækkelige – for eksempel perfekt kvalitetssikring. Det relevante spørgsmål er, hvor den nedre grænse for tilstrækkelig kvalitetssikring findes – altså præcis hvilket niveau af kvalitetssikring, som er lige akkurat nok.

Urimeligt-krav-indvendingen

Vi bør ikke stille højere krav til kvalitetssikring, end eksisterende metoder kan levere, i brugsscenarier hvor fejl har meget små konsekvenser.

En anden mulig indvending er, at positionen synes at stille urimeligt høje krav, især i brugsscenarier hvor konsekvenserne af fejl er meget beskedne. Nogle kunne eksempelvis mene, at det ikke er afgørende at kvalitetssikre GenAI til tekstredigering, fordi en fejl typisk vil bestå i at GenAI overser stavfejl eller grammatiske fejl. Selv hvis

eksisterende metoder til kvalitetssikring er relativt upræcise, så kan det forekomme urimeligt krævende, at afvise anvendelse af GenAI, med den begrundelse, at den kun fanger nogle sproglige fejl, og at vi ikke ved præcist hvor mange.

Standard

En anden mulig position er, at offentlige myndigheder bør anvende et standardiseret sæt tests til at foretage kvalitetssikring. Det vil nok forekomme nogle at være et rimeligt synspunkt, blandt andet fordi vi ofte anvender standardiserede tests på andre teknologier. Standardiserede test har også den fordel, at man derved får et ensartet billede af risici på tværs af brugsscenarier, samtidig med, at det kan være lettere at kvalitetssikre, fordi man kan oparbejde rutiner i at udføre de pågældende test.

Standard

Offentlige myndigheder bør anvende et standardiseret sæt tests til kvalitetssikring af GenAI.

Positionen må gøres mere konkret, ved at pege på hvad den relevante standard er. Her kan man forestille sig både mere ambitiøse og mere moderate versioner. En ambitiøs version af positionen kunne være, at offentlige myndigheder til enhver tid bør anvende den kvalitetssikring, som giver så præcist et billede som muligt af systemets risiko for fejl. Men man kan også forestille sig en moderat version af positionen, hvor man hævder at offentlige myndigheder bør anvende et standardiseret sæt, som balancerer hensyn til kvalitetssikring med hensynet til at begrænse omkostningerne til kvalitetssikring.

En indvending mod positionen kunne være, at den ved at kræve én standard for alle brugsscenarier risikerer at være både for meget og for lidt. I nogle tilfælde, kunne kritikerne hævde, kan det føre til

For-meget-og-for-lidt-indvendingen

En standard for kvalitetssikring, vil være for høj i nogle brugsscenarier og/eller for lav i andre.

at man skyder gråspurve med kanoner, ved at anvende meget omfattende kvalitetssikring på GenAI i et brugsscenarie, hvor det dataetisk set ikke er synderligt vigtigt, hvad risikoen for fejl er, for eksempel fordi konsekvenserne af fejl er trivielle. I andre tilfælde kunne man komme til at sætte barren for lavt, måske fordi GenAI anvendes på en måde, hvor det er afgørende at systemet ikke begår fejl.

Tilpasset

En tredje mulig position er, at kvalitetssikring bør variere fra brugsscenarie til brugsscenarie, afhængigt af hvilke risici systemet har, og de omkostninger, som forskellige kvalitetsvurderinger medfører. Denne position kan nok forekomme troværdig, hvis man antager det synspunkt, som den foregående indvending hvilede på, at der kan være forskelle på brugsscenarier, som påvirker hvilke krav vi bør stille til kvalitetssikring. Man kan forestille sig, at visse anvendelser af GenAI har en risikoprofil som betyder, at det kun er

Tilpasset

GenAI bør kvalitetssikres med et sæt af tests, som svarer til vigtigheden af at vurdere systemets risiko præcist og pålideligt, således at systemer hvor fejl har mere alvorlige konsekvenser vurderes mere grundigt, og systemer hvor fejl har mindre alvorlige konsekvenser vurderes mindre grundigt.

nødvendigt, at lave meget beskeden kvalitetssikring, mens andre anvendelser gør det nødvendigt, at foretage meget grundig kvalitetssikring. Det er værd at bemærke, at der er mange mulige måder at gøre positionen mere konkret og præcis. Nogle vil nok have lyst til at stille større krav til kvalitetssikring, mens andre vil stille mindre krav til kvalitetssikring. En mulig variant af positionen kan således minde om den første position, idet man kunne hævde, at risikoen for fejl i visse brugsscenarier er så alvorlig, at eksisterende metoder til kvalitetssikring er utilstrækkelige, og man derfor bør undlade at anvende GenAI på disse måder.

Ligesom ved de øvrige positioner, er man nødt til at specificere, hvad det nødvendige niveau af kvalitetssikring er, for forskellige brugsscenarier, og forklare hvorfor det er netop dette niveau, som dataetisk set er nødvendigt. De afvejninger, som man i den forbindelse vil foretage, kan antageligt minde en del, om de afvejninger, man foretager i forhold til det første kernespørgsmål, hvis man mener at den dataetisk acceptable risiko for fejl afhænger af en samlet vurdering af fordele og ulemper ved at anvende GenAI. En indvending i den forbindelse var, at den samlede vurdering ville være for vanskelig at operationalisere. En lignende indvending kunne her hævde, at et krav til kvalitetssikring som betyder, at man skal foretage komplekse dataetiske afvejninger i hvert enkelt brugsscenarie, er for vanskeligt at anvende i praksis.

Også-for-krævende-indvendingen

Vi bør ikke stille krav til kvalitetssikring, som forudsætter dataetisk afvejning, der er for vanskelig at operationalisere.



6. Risiko for fejl i offentlige myndigheders anvendelse af generativ AI: Dataetisk Råds position og anbefalinger

Generativ AI er en banebrydende teknologi, som kan anvendes i mange forskellige brugsscenarier. Teknologien har blandt andet potentiale til at forbedre effektivitet og kvalitet i den offentlige sektor. Dataetisk Råd støtter udforskning af disse potentialer i brugsscenarier, hvor offentlige myndigheder kan realisere gevinster for borgere og medarbejdere.

Offentlige myndigheders anvendelse af GenAI rejser imidlertid også en række dataetiske udfordringer, hvor udvikling og anvendelse af GenAI kan have negative konsekvenser for mennesker, miljø og samfund. Det gælder for eksempel risiko for fejl, bias og læk af følsomme data, miljøpåvirkning ved udvikling og anvendelse af GenAI, krænkelse af intellektuel ejendomsret ved træning af GenAI, og styrkelse af tech-oligopoler (se kapitel 2). Mange af disse dataetiske udfordringer er underbelyste. Eksisterende retningslinjer behandler således kun i begrænset omfang den centrale dataetiske udfordring, at GenAI-systemer har risiko for at lave fejl, og spørgsmålet om hvordan man kan kvalitetssikre GenAI, ved at teste risikoen for fejl (se kapitel 4).



Det er **Dataetisk Råds opfattelse**, at det er afgørende, at vi som samfund forholder os til de dataetiske udfordringer, som udvikling og anvendelse af GenAI rejser.

”Det er afgørende, at vi som samfund forholder os til de dataetiske udfordringer, som udvikling og anvendelse af GenAI rejser.”

Offentlige myndigheder har en særlig rolle, når de anvender ny digital teknologi som GenAI. Det skyldes myndigheders ansvar for at formidle korrekt information, og for at træffe korrekte beslutninger i sager der berører borgere, både af hensyn til den enkelte borger, og af hensyn til den værdifulde tillid, som danskere har til offentlige myndigheder.

Det er **Dataetisk Råds opfattelse**, at de offentlige myndigheder, som er med til at udbrede anvendelsen af GenAI, har et særligt ansvar for at kombinere denne indsats med et øget fokus på hvordan myndigheder kan tackle dataetiske udfordringer, og sikre dataetisk anvendelse af GenAI.

”Offentlige myndigheder, som anvender GenAI, har et særligt ansvar for at tackle de dataetiske udfordringer.”

6.1. Dataetiske overvejelser før anvendelse af generativ AI

Mange offentlige myndigheder i Danmark har været hurtige til at igangsætte udforskning og test af, hvordan myndigheder kan anvende GenAI (se kapitel 3). Den offentlige sektor i Danmark er digitaliseret, og de mange nye projekter sigter mod på forskellige måder at effektivisere myndighedernes arbejde, og forbedre den offentlige service ved at bruge GenAI.

Dataetisk Råd støtter at offentlige myndigheder udforsker og afprøver hvordan teknologien kan anvendes. Det skyldes at rådet vurderer, at teknologien har potentiale til at realisere vigtige gevinster for borgerne, og fordi det er rådets opfattelse, at dette arbejde er en uundgåelig del af en nødvendig læreproces i den offentlige sektor. Kun ved at arbejde med teknologien får danske myndigheder mulighed for at afprøve brugsscenarier, oparbejde tekniske kompetencer, udvikle nødvendige administrative rammer, og opdage dataetiske dilemmaer og grænser for anvendelse.

Offentlige myndigheder er imidlertid særligt sårbare under denne læreproces, hvor de oparbejder tekniske kompetencer, udvikler nødvendige administrative rammer, og etablerer de dataetiske grænser for anvendelse. Udforskning og læring bør foretages med blik for både denne sårbarhed og de forskellige risici, som findes i forskellige brugsscenarier. Når en offentlig myndighed går i gang med at udforske og anvende GenAI, bør myndigheden fokusere på de brugsscenarier, hvor man kan udforske og afprøve teknologien med lavest mulig risiko for borgere og medarbejdere.

En vigtig metode til at reducere sårbarhed og accelerere læring, er inddragelse af interessenter i udviklingen af GenAI-systemer, og i evalueringen af systemer som er testet eller sat i drift. Offentlige myndigheder, som ønsker at udforske og anvende GenAI, kan med fordel konsultere berørte borgere og medarbejdere, både individuelt og gennem inddragelse af relevante faglige og civilsamfundsorganisationer.

Dataetisk Råd anbefaler, at offentlige myndigheder begynder eventuel anvendelse af GenAI, ved at anvende GenAI-systemer, som har en lav risikoprofil for fejl. Hvis myndigheder ønsker at udvide til andre brugsscenarier, hvor GenAI har en højere risikoprofil, bør denne udvidelse ske gradvist, i takt med at myndigheden opbygger erfaringer, kompetencer og afprøvede retningslinjer for anvendelse. I alle tilfælde, bør myndigheder styrke læreprocessen og reducere risiko for fejl, ved at inddrage interessenter i udvikling og evaluering af systemer.

Anbefaling 1

Offentlige myndigheder bør begynde eventuel udforskning og anvendelse af GenAI med brugsscenarier, hvor anvendelsen har lav risikoprofil, og inddrage interessenter i udvikling og evaluering af systemer.

Risikoen for fejl er den måske mest alvorlige dataetiske udfordring ved offentlige myndigheders brug af GenAI. GenAI-systemer kan fejle ved at skabe faktisk forkert indhold, ved at udelade relevant indhold, og ved at skabe indhold, der er uegnet til formålet, for eksempel fordi det er svært at forstå (se kapitel 5). Et centralt dataetisk princip er fordringen om, at databehandling bør gøre gavn, ikke skade. Når et GenAI-system fejler, kan det føre til, at myndigheder eller borgere træffer fejlagtige beslutninger, der i sidste ende skader borgere. Der er således stærke dataetiske grunde til, at offentlige myndigheder ikke bør tillade fejl i GenAI-systemer.

Alle systemer har imidlertid risiko for at fejle. Hvis man stillede krav om, at GenAI-systemer skulle være ufejlbarlige, så ville man derfor i praksis gøre det umuligt at anvende

GenAI. Et sådant krav ville også sætte tærsklen for GenAI-systemer langt højere, end de krav vi stiller til for eksempel menneskelige systemer.

Omvendt ville ukritisk anvendelse af GenAI, uden hensyn til systemets risiko for fejl, være dataetisk uansvarligt. Vi forventer med rette, at offentlige myndigheder gør sig store anstrengelser for at undgå fejl.

Imellem disse yderpositioner findes forskellige syn på, hvilke krav man bør stille til GenAI-systemers risikoprofil. Det er Dataetisk Råds opfattelse, at offentlige myndigheder bør vurdere om et GenAI-systems risikoprofil er for høj, ved at sammenligne med de risici relevante alternativer medfører. Den eksisterende praksis før myndigheden indfører GenAI-systemet vil altid være et relevant alternativ, som man bør sammenligne med. Denne status quo vil ofte være en praksis hvor mennesker løser den opgave, som GenAI skal udføre. De forskellige måder, man teknisk kunne ændre på et GenAI-system, eller på implementeringen af systemet ved myndigheden, vil også være relevante alternativer. Det bør i denne forbindelse ikke være en legitim begrundelse, at GenAI vil være en forbedring relativt til status quo, hvis dette skyldes, at kvaliteten i myndighedens arbejde er lavere end vi som samfund ville eller burde acceptere. I denne situation vil det være relevant at sammenligne med alternative måder, at forbedre kvaliteten i myndighedens arbejde.

Dataetisk Råd anbefaler, at offentlige myndigheder skelner mellem GenAI-systemer med højere og lavere risikoprofil, og alene anvender GenAI, når myndigheden har indført mitigerende tiltag for at begrænse risikoen for fejl, der er tilstrækkelige i forhold til systemets risikoprofil. Konkret bør myndigheder stille stærkere krav til mitigering af risiko for fejl, jo højere systemets risikoprofil er.

Anbefaling 2

Offentlige myndigheder bør skelne mellem GenAI-systemer med høj og lav risikoprofil, og kun bruge GenAI, når myndigheden har indført tilstrækkelige mitigerende tiltag til at begrænse risikoen for fejl.

6.2. Dataetiske tiltag ved anvendelse af generativ AI

Risikoen for fejl er en central dataetisk udfordring, når offentlige myndigheder anvender GenAI, men risikoen varierer afhængigt af hvordan GenAI anvendes. I nogle brugsscenarier kan risikoen for fejl være en begrænset udfordring, i andre brugsscenarier kan risikoen for fejl være en væsentlig udfordring. Offentlige myndigheder bør derfor foretage en systematisk vurdering af udfordringen med risiko for fejl i hvert enkelt brugsscenarie, hvor myndigheden anvender GenAI.

Myndigheders systematiske vurdering af risikoen for fejl drejer sig i sidste ende om at vurdere risikoen for, at anvendelse af GenAI får negativ betydning for borgere og medarbejdere. Når offentlige myndigheder vurderer risikoen for fejl, bør de derfor tage hensyn til, at GenAI-systemer anvendes i en social og administrativ kontekst. Den dataetiske udfordring med risiko for fejl kan være større eller mindre afhængigt af flere faktorer. Det spiller en rolle hvor mange og hvilke fejl systemet laver, hvor mange fejl myndigheden mitigerer efter systemet har genereret indhold, hvilken indflydelse ikke-mitigerede fejl har på menneskelige beslutninger, og hvilke beslutninger systemet påvirker. Tilsammen bestemmer disse forhold et GenAI-systems risikoprofil med hensyn til fejl (se kapitel 5).

Dataetisk Råd anbefaler, at offentlige myndigheder vurderer GenAI-systemers samlede risikoprofil for fejl som en funktion af mængden og karakteren af fejl, antallet af fejl som mitigeres, systemets indvirkning på beslutninger, samt beslutningernes konsekvenser for berørte borgere og medarbejdere.

Anbefaling 3

Offentlige myndigheder bør systematisk vurdere GenAI-systemers samlede risikoprofil for fejl i hvert enkelt brugsscenarie.

En central del af vurderingen af et GenAI-systems risikoprofil er kvalitetssikring, som tester systemet for at finde ud af, hvor mange og hvilke fejl systemet laver. Det kaldes i faglitteraturen for "modelvalidering". Kvalitetssikring af GenAI rejser en selvstændig dataetisk problemstilling, fordi det ofte er langt vanskeligere at udføre sådanne tests på GenAI-systemer end på andre former for kunstig intelligens. Risikoen kan variere på tværs af brugsscenarier, og eksisterende tekniske metoder har vanskeligt ved at måle risikoen præcist og pålideligt. Hvis en myndighed vil have et præcist billede af et GenAI-systems risiko for fejl, kan det derfor være nødvendigt, at gennemføre res-

sourcetunge manuelle tests, hvor mennesker kigger på mange eksempler på AI-skabt indhold, og vurderer kvaliteten (se kapitel 5).

Dataetisk Råd anbefaler, at offentlige myndigheder tilpasser graden af kvalitetssikring til systemets øvrige risikoprofil i det specifikke brugsscenarie. Det betyder, at myndigheder bør udføre mere omfattende test af et GenAI-systems risiko for at fejle, jo færre fejl myndigheden mitigerer, jo større indflydelse fejl har på beslutninger, og jo vigtigere de relevante beslutninger er.

Anbefaling 4

Offentlige myndigheder bør gennemføre mere omfattende kvalitetssikring af et GenAI-system, jo højere systemets øvrige risikoprofil er.

Mange dataetiske retningslinjer for kunstig intelligens anbefaler, at organisationen håndterer risikoen for fejl, ved at lade mennesker kontrollere kvaliteten af GenAI-indhold (se kapitel 4). En myndighed kan for eksempel pålægge medarbejdere ansvaret for at identificere og korrigere eventuelle fejl i de resumeer af sager, som myndighedens GenAI-system skaber.

Menneskelig kontrol af fejl har imidlertid indbyggede svagheder, fordi sådan kontrol kan være meget ressourcekrævende, fordi det kan være vanskeligt for mennesker at vurdere kvaliteten af GenAI-indhold, og fordi man ikke kan forudsætte, at borgere kan kontrollere for fejl, når GenAI bruges borgerrettet.

Anbefaling 5

Offentlige myndigheder bør oplyse, hvordan de anvender menneskelig kvalitetskontrol af GenAI, og tildele relevante medarbejdere de nødvendige ressourcer.

Dataetisk Råd anbefaler, at offentlige myndigheder gør det klart for både borgere og medarbejdere hvornår og hvordan medarbejdere udfører menneskelig kontrol med fejl, samt at myndigheden sikrer, at de pågældende medarbejdere har de nødvendige ressourcer til at udføre effektiv kontrol, i form af eksempelvis politikker, kompetenceudvikling, træning og tid.

Fordi offentlige myndigheders anvendelse af GenAI-systemer ikke kan være fejlfri, er det vigtigt, at myndigheder som anvender GenAI tager relevante forholdsregler. Disse forholdsregler bør sigte mod, at begrænse de negative konsekvenser, som fejl kan

have for medarbejdere og borgere. De kan inkludere procedurer for at identificere og verificere fejl i indhold efter at myndigheden har anvendt GenAI-indhold, praksisser for at oplyse berørte borgere, kontakt- og klageadgang for borgere, som mener at der har eller kan have været fejl, lagring af relevante materialer, for at dokumentere eventuelle fejl, samt procedurer for at rette fejl og kompensere berørte borgere.

Dataetisk Råd anbefaler, at offentlige myndigheder som anvender GenAI indfører forholdsregler der sikrer, at det er let for borgere og medarbejdere at identificere fejl, at dokumentere fejl, at oplyse om fejl, og at korrigere fejl, herunder udbedre konsekvenserne af fejl. Sådanne forholdsregler bør udvikles og indføres inden det relevante GenAI-system tages i brug.

Anbefaling 6

Offentlige myndigheder bør indføre nødvendige forholdsregler til at identificere, dokumentere, oplyse og korrigere fejl inden, de tager et GenAI-system i brug.

6.3. Politiske initiativer for offentlige myndigheders anvendelse af GenAI

Dataetisk Råd forventer, at kunstig intelligens kommer til at spille en hastigt voksende rolle i den offentlige sektor i Danmark. Det er imidlertid rådets opfattelse, at udbredelse og etisk anvendelse af GenAI i den offentlige sektor hæmmes af, at offentlige myndigheder på visse områder skal navigere i et utrygt rum af muligheder, hvor det er uklart, hvordan de kan og bør tackle dataetiske udfordringer. Mange af de dataetiske udfordringer, som den enkelte myndighed møder i arbejdet med GenAI, er udfordringer, som også gør sig gældende for beslægtede brugsscenarier ved andre myndigheder. Der findes også samarbejde mellem myndigheder om at håndtere udfordringerne i formelle og uformelle netværk. I mange tilfælde er ansvaret for, at forholde sig til de dataetiske udfordringer, ikke desto mindre overladt til den enkelte myndighed.

Konsekvensen af at myndigheder individuelt skal tackle dataetiske udfordringer er, at den enkelte myndighed unødigt bruger ressourcer på at foretage vanskelige dataetiske vurderinger, som kunne gennemføres mere effektivt og med større effekt kollektivt. Samtidig kan manglen på fælles retningslinjer føre til, at nogle myndigheder tøver, hvor de med fordel kunne gå frem, mens andre risikerer at gå for langt.

I lyset af ulemperne ved, at overlade ansvaret for at tackle de dataetiske udfordringer til den individuelle myndighed, kan den offentlige sektor med fordel i højere grad samarbejde om at udveksle erfaringer og udvikle fælles praksisser, retningslinjer og standarder for, hvordan man håndterer dataetiske udfordringer i anvendelsen af GenAI.

Dataetisk Råd anbefaler, at Digitaliseringsministeren tager initiativ til udvikling af fællesoffentlige retningslinjer for håndtering af udfordringen med risiko for fejl ved offentlige myndigheders anvendelse af GenAI. Disse retningslinjer bør præcisere hvordan myndigheder skal definere tærskler for en acceptabel risikoprofil, og fastsætte hvilke skridt myndigheder skal tage, for at kvalitetssikre GenAI-systemer.

Anbefaling 7

Regeringen bør udvikle fælles retningslinjer for offentlige myndigheders anvendelse af GenAI, som slår fast, hvordan myndigheder skal håndtere risikoen for fejl.

Offentlige myndigheders anvendelse af GenAI rejser en række forskellige dataetiske udfordringer. Nogle af disse udfordringer knytter sig til den enkelte myndigheds anvendelse i et konkret brugsscenarie, og nogle knytter sig til den samlede offentlige sektor. Mange af udfordringerne har vi som samfund kun akkurat taget hul på at forholde os til, og finde løsninger på. Disse løsninger vil i de kommende år være et centralt emne for vores fælles demokratiske debat.

Den fælles samtale om, hvordan vi ønsker at bruge GenAI i den offentlige sektor, udfordres imidlertid af manglende kendskab til de måder GenAI anvendes og kan anvendes. Myndigheders mulighed for at udveksle erfaringer og dataetiske refleksioner om GenAI vanskeliggøres også af, at der ikke findes et samlet overblik over, hvordan offentlige myndigheder i Danmark anvender GenAI i særdeleshed, og kunstig intelligens mere generelt.

Der findes internationalt en række erfaringer med registre for AI, blandt andet i Helsinki og Amsterdam, og i det såkaldte "Eurocities Digital Forum Lab"-samarbejde. En række danske aktører har også taget prisværdige initiativer til at skabe overblik over offentlige myndigheders anvendelse af kunstig intelligens. Det gælder Datatilsynet, KL, og AI & Human Rights-projektet ved Jura på Københavns Universitet (se kapitel 3). Dataetisk Råds opfattelse er ikke desto mindre, at der er behov for at placere et centralt ansvar for denne opgave. Det er Dataetisk Råds opfattelse, at det vil være en væsentlig fordel for både den offentlige sektor, og for den offentlige, demokratiske debat, at skabe et velfunderet, løbende opdateret overblik over, hvordan offentlige myndigheder anvender

der GenAI. Et sådant overblik bør som minimum indeholde informationer om hvilke data systemet anvender, systemernes tekniske specifikationer, herunder hvordan systemet eventuelt er fintunet, og data-arkitekturen i den implementerede løsning, samt om hvordan systemerne integreres i myndighedens arbejde.

Dataetisk Råd anbefaler, at Digitaliseringsministeren tager initiativ til oprettelse af et nationalt offentligt register over GenAI-systemer, som anvendes af offentlige myndigheder. Registreret bør hente inspiration fra både danske og internationale erfaringer med sådanne registre, og forankres ved en egnet dansk myndighed. Etablering kan med fordel gøres til en opgave for regeringens task-force for kunstig intelligens.

Anbefaling 8

Regeringen bør oprette et nationalt, offentligt register over GenAI-systemer, som anvendes af offentlige myndigheder.

Dataetisk Råds anbefalinger

Det er afgørende, at vi som samfund forholder os til de dataetiske udfordringer, som udvikling og anvendelse af GenAI rejser. Offentlige myndigheder, som anvender GenAI, har et særligt ansvar for at tackle disse dataetiske udfordringer.

- 1. Offentlige myndigheder** bør begynde eventuel udforskning og anvendelse af GenAI med brugsscenarier, hvor anvendelsen har lav risikoprofil, og inddrage interessenter i udvikling og evaluering af systemer.
- 2. Offentlige myndigheder** bør skelne mellem GenAI-systemer med høj og lav risikoprofil, og kun bruge GenAI, når myndigheden har indført tilstrækkelige mitigerende tiltag til at begrænse risikoen for fejl.
- 3. Offentlige myndigheder** bør systematisk vurdere GenAI-systemers samlede risikoprofil for fejl i hvert enkelt brugsscenarie.
- 4. Offentlige myndigheder** bør gennemføre mere omfattende kvalitets-sikring af et GenAI-system, jo højere systemets risikoprofil er.
- 5. Offentlige myndigheder** bør oplyse, hvordan de anvender menneskelig kvalitetskontrol af GenAI, og tildele relevante medarbejdere de nødvendige ressourcer.
- 6. Offentlige myndigheder** bør indføre nødvendige forholdsregler til at identificere, dokumentere, oplyse og korrigere fejl, inden de tager et GenAI-system i brug.
- 7. Regeringen** bør udvikle fælles retningslinjer for offentlige myndighedsers anvendelse af GenAI, som slår fast hvordan myndigheder skal håndtere risikoen for fejl.
- 8. Regeringen** bør oprette et nationalt, offentligt register over GenAI-systemer, som anvendes af offentlige myndigheder.

Litteratur

Alammar, J. (2018). The Illustrated Transformer. <https://jalammar.github.io/illustrated-transformer/>.

Arkoudas, K. (2023). "ChatGPT is no Stochastic Parrot. But it also Claims that 1 is Greater than 1." *Philosophy & Technology* 36(3): 54 DOI: 10.1007/s13347-023-00619-6.

Ayyamperumal, S. G. & Ge, L. (2024). "Current state of LLM Risks and AI Guardrails." arXiv preprint arXiv:2406.12934.

Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?? *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*.

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J. & Caliskan, A. (2023). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A. & Brunskill, E. (2021). "On the opportunities and risks of foundation models." arXiv preprint arXiv:2108.07258.

Bostrom, N. (2011). "Information hazards: A typology of potential harms from knowledge." *Review of Contemporary Philosophy*(10): 44-79.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press.

Bringsjord, S. & Govindarajulu, N. S. (2018). Artificial Intelligence. *Stanford Encyclopedia of Philosophy*. Edward N. Zalta. <https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=artificial-intelligence>.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y. & Lundberg, S. (2023). "Sparks of artificial general intelligence: Early experiments with gpt-4." arXiv preprint arXiv:2303.12712.

Buchanan, B., Lohn, A., Musser, M. & Sedova, K. (2021). "Truth, lies, and automation." *Center for Security and Emerging technology* 1(1): 2.

Burton, J. W., Stein, M.-K. & Jensen, T. B. (2020). "A systematic review of algorithm aversion in augmented decision making." *33*(2): 220-239 DOI: 10.1002/bdm.2155.

Chesterman, S. (2024). "Good models borrow, great models steal: intellectual property rights and generative AI." *Policy and Society* DOI: 10.1093/polsoc/puae006.

- Conitzer, V. & Leben, D. (2024). How ChatGPT has been prompted to respect safety fairness, and copyright. AI Ethics at Oxford Blog. <https://www.oxford-aiethics.ox.ac.uk/blog/how-chatgpt-has-been-prompted-respect-safety-fairness-and-copyright>.
- Cummings, M. "Automation Bias in Intelligent Time Critical Decision Support Systems." AIAA 1st Intelligent Systems Technical Conference.
- Den uafhængige ekspertgruppe på højt niveau om kunstig intelligens (2018). Ethiske retningslinjer for pålidelig kunstig intelligens, Europa-Kommissionen. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60420.
- Dietvorst, B. J., Simmons, J. P. & Massey, C. (2015). "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err." *Journal of Experimental Psychology: General* 144(1): 114-126.
- Dong, Y., Mu, R., Jin, G., Qi, Y., Hu, J., Zhao, X., Meng, J., Ruan, W. & Huang, X. (2024). "Building guardrails for large language models." arXiv preprint arXiv:2402.01822.
- Eloundou, T., Manning, S., Mishkin, P. & Rock, D. (2023). "Gpts are gpts: An early look at the labor market impact potential of large language models." arXiv preprint arXiv:2303.10130.
- Europol (2023). ChatGPT – The impact of Large Language Models on Law Enforcement Luxembourg, Europol DOI: 10.2813/255453.
- Fang, A. & Perkins, J. M. (2024). "Large language models (LLMs): Risks and policy implications." MIT Science Policy Review.
- Felten, E., Raj, M. & Seamans, R. (2023). "How will language models like chatgpt affect occupations and industries?" arXiv preprint arXiv:2303.01157.
- Flynn, A., Powell, A., Scott, A. J. & Cama, E. (2021). "Deepfakes and Digitally Altered Imagery Abuse: A Cross-Country Exploration of an Emerging form of Image-Based Sexual Abuse." *The British Journal of Criminology* 62(6): 1341-1358 DOI: 10.1093/bjc/azab111.
- Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Deroncourt, F., Yu, T., Zhang, R. & Ahmed, N. K. (2024). "Bias and Fairness in Large Language Models: A Survey." *Computational Linguistics* 50(3): 1097-1179 DOI: 10.1162/coli_a_00524.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M. & Wang, H. (2023). "Retrieval-augmented generation for large language models: A survey." arXiv preprint arXiv:2312.10997.

- George, A. (2023) "Visualizing the size of Large Language Models." Medium. <https://medium.com/@georgeanil/visualizing-size-of-large-language-models-ec576caa5557>.
- Goddard, K., Roudsari, A. & Wyatt, J. C. (2011). "Automation bias: a systematic review of frequency, effect mediators, and mitigators." *Journal of the American Medical Informatics Association* 19(1): 121-127 DOI: 10.1136/amiajnl-2011-000089 %J *Journal of the American Medical Informatics Association*.
- Gupta, S., Ranjan, R. & Singh, S. N. (2024). "A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions." arXiv preprint arXiv:2410.12837.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. & Steinhardt, J. (2020). "Measuring massive multitask language understanding." arXiv preprint arXiv:2009.03300.
- Implement (2023). *Sprogmodeller i Danmark - En analyse af mulige strategiske valg og scenarier*. <https://www.kl.dk/media/gyopkwnu/sprogmodeller-i-danmark.pdf>.
- Janjeva, A., Harris, A., Mercer, S., Kasprzyk, A. & Gausen, A. (2023). *The Rapid Rise of Generative AI: Assessing risks to safety and security*. London, UK, The Alan Turing Institute. https://cetas.turing.ac.uk/sites/default/files/2023-12/cetas_research_report_-_the_rapid_rise_of_generative_ai_-_2023.pdf.
- Jones, E., Dunlop, C. & Ghani, B. (2023). *What is a foundation model?* London, UK, Ada Lovelace Institute <https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/>.
- Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., Hopkins, A., Bankston, K., Biderman, S. & Bogen, M. (2024). "On the Societal Impact of Open Foundation Models." arXiv preprint arXiv:2403.07918.
- Kapoor, S. & Narayanan, A. (2023). *Quantifying ChatGPT's Gender Bias*. AI Snake Oil. <https://www.aisnakeoil.com/p/quantifying-chatgpts-gender-bias>.
- KL & Digitaliseringsstyrelsen (2024). *Inspirationskatalog - Syv eksempler med generativ AI i den offentlige sektor*. https://digst.dk/media/31743/inspirationskatalog-syv-eksempler-med-generativ-ai-i-den-offentlige-sektor-webtilgaengelig_final.pdf.
- Laban, P., Schnabel, T., Bennett, P. N. & Hearst, M. A. (2022). "SummaC: Re-visiting NLI-based models for inconsistency detection in summarization." *Transactions of the Association for Computational Linguistics* 10: 163-177.
- Lee, J. D. & See, K. A. (2004). "Trust in automation: Designing for appropriate reliance." *Human Factors* 46(1): 50-80.
- Li, P., Yang, J., Islam, M. A. & Ren, S. (2023). "Making ai less" thirsty": Uncovering and addressing the secret water footprint of ai models." arXiv preprint arXiv:2304.03271.

- Lin, S., Hilton, J. & Evans, O. (2021). "Truthfulqa: Measuring how models mimic human falsehoods." arXiv preprint arXiv:2109.07958.
- Lorenz, P., Perset, K. & Berryhill, J. (2023). Initial Policy Considerations for Generative Artificial Intelligence. Paris, France, OECD. <https://doi.org/10.1787/fae2d1e6-en>.
- Lucchi, N. (2023). "ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems." *European Journal of Risk Regulation*: 1-23 DOI: 10.1017/err.2023.59.
- Luo, G., Fan, W., Li, M., He, Y., Yang, Y. & Bao, F. (2024). On the Intractability to Synthesize Factual Inconsistencies in Summarization. Findings of the Association for Computational Linguistics: EACL 2024.
- Marchal, N., Xu, R., Elasmr, R., Gabriel, I., Goldberg, B. & Isaac, W. (2024). "Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data." arXiv preprint arXiv:2406.13843.
- Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M. & Cerf, M. (2024). "The potential of generative AI for personalized persuasion at scale." *Scientific Reports* 14(1): 4692 DOI: 10.1038/s41598-024-53755-0.
- Maynez, J., Narayan, S., Bohnet, B. & McDonald, R. (2020). "On faithfulness and factuality in abstractive summarization." arXiv preprint arXiv:2005.00661.
- McCarthy, J. (2007). What is artificial intelligence? <http://cse.unl.edu/~choueiry/S09-476-876/Documents/whatisai.pdf>.
- Myrstad, F. L.-H. (2023). Ghost in the Machine – Addressing the Consumer Harms of Generative AI. Oslo, Norway, Forbrukerrådet Norge.
- Naitali, A., Ridouani, M., Salahdine, F. & Kaabouch, N. (2023). "Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions." *Computers* 12(10): 216. <https://www.mdpi.com/2073-431X/12/10/216>.
- Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F. & Lee, K. (2023). "Scalable extraction of training data from (production) language models." arXiv preprint arXiv:2311.17035.
- Nguyen, H., Chen, H., Pobbathi, L. & Ding, J. (2024). "A Comparative Study of Quality Evaluation Methods for Text Summarization." arXiv preprint arXiv:2407.00747.
- Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V. & Nguyen, C. M. (2022). "Deep learning for deepfakes creation and detection: A survey." *Computer Vision and Image Understanding* 223: 103525 DOI: <https://doi.org/10.1016/j.cviu.2022.103525>.
- Nicoletti, L. & Bass, D. (2023) "Humans are biased. Generative AI is even worse." Bloomberg. <https://www.bloomberg.com/graphics/2023-generative-ai-bias/>.

- Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J. & Zhang, T. (2023). "Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models." arXiv preprint arXiv:2401.00396.
- Novelli, C., Taddeo, M. & Floridi, L. (2024). "Accountability in artificial intelligence: what it is and how it works." *Ai & Society* 39(4): 1871-1882.
- Pagnoni, A., Balachandran, V. & Tsvetkov, Y. (2021). "Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics." arXiv preprint arXiv:2104.13346.
- Passi, S. & Vorvoreanu, M. (2022). Overreliance on AI: Literature Review, Microsoft. <https://www.microsoft.com/en-us/research/uploads/prod/2022/06/Aether-Overreliance-on-AI-Review-Final-6.21.22.pdf>.
- Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M. & Dean, J. (2021). "Carbon emissions and large neural network training." arXiv preprint arXiv:2104.10350.
- Prahl, A. & Van Swol, L. (2017). "Understanding algorithm aversion: When is advice from automation discounted?" *Journal of Forecasting* 36(6): 691-702 DOI: 10.1002/for.2464.
- Santoni de Sio, F. (2024). "Artificial Intelligence and the Future of Work: Mapping the Ethical Issues." *The Journal of Ethics* 28(3): 407-427 DOI: 10.1007/s10892-024-09493-6.
- Simchon, A., Edwards, M. & Lewandowsky, S. (2024). "The persuasive effects of political microtargeting in the age of generative artificial intelligence." *PNAS Nexus* 3(2) DOI: 10.1093/pnasnexus/pgae035.
- Tang, L., Goyal, T., Fabbri, A. R., Laban, P., Xu, J., Yavuz, S., Kryściński, W., Rousseau, J. F. & Durrett, G. (2022). "Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors." arXiv preprint arXiv:2205.12854.
- Tang, L., Laban, P. & Durrett, G. (2024). "MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents." arXiv preprint arXiv:2404.10774.
- UK, O. (2024). Deepfake Defences - Mitigating the Harms of Deceptive Deepfakes. <https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/deepfake-defences/>.
- Vaswani, A. (2017). "Attention is all you need." *Advances in Neural Information Processing Systems*.
- Walker, S. M. (2023) "The Next Era of AI: Inside the Breakthrough GPT-4 Model." Hackernoon. https://hackernoon.com/the-next-era-of-ai-inside-the-breakthrough-gpt-4-model?utm_source=chatgpt.com.
- Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O. & Bowman, S. (2019). "Superglue: A stickier benchmark for general-purpose language understanding systems." *Advances in neural information processing systems* 32.

- Wang, P. (2019). "On defining artificial intelligence." *Journal of Artificial General Intelligence* 10(2): 1-37.
- Webb, M. (2019). "The impact of artificial intelligence on the labor market." Available at SSRN 3482150.
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B. & Kasirzadeh, A. (2021). "Ethical and social risks of harm from language models." arXiv preprint arXiv:2112.04359.
- Xu, D., Fan, S. & Kankanhalli, M. (2023). Combating misinformation in the era of generative AI models. *Proceedings of the 31st ACM International Conference on Multimedia*.
- Yan, B., Li, K., Xu, M., Dong, Y., Zhang, Y., Ren, Z. & Cheng, X. (2024). "On protecting the data privacy of large language models (LLMs): A survey." arXiv preprint arXiv:2403.05156.
- Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z. & Zhang, Y. (2024). "A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly." *High-Confidence Computing* 4(2): 100211 DOI: <https://doi.org/10.1016/j.hcc.2024.100211>.
- Zerilli, J., Knott, A., Maclaurin, J. & Gavaghan, C. (2019). "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?" *Philosophy & Technology* 32: 661-683 DOI: 10.1007/s13347-018-0330-6.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. (2019). "Bertscore: Evaluating text generation with bert." arXiv preprint arXiv:1904.09675.
- Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D. & Du, M. (2024). "Explainability for Large Language Models: A Survey." *ACM Trans. Intell. Syst. Technol.* 15(2): Article 20 DOI: 10.1145/3639372.



Dataetisk Råd 2025

Dataetisk Råd er en offentlig myndighed ved Nationalt Center for Etik, som rådgiver og skaber debat om etisk udvikling og anvendelse af digitale løsninger, data og kunstig intelligens.

København, 8. Januar 2025