Digitaliserings-
styrelsen

Notat

11. marts 2024

Guide til offentlige myndigheder om ansvarlig anvendelse af generativ kunstig intelligens

Version nr.	Dato	Udarbejdet / revideret af	Ændringer/ bemærkninger
1	Jan. 2024	Digitaliseringsstyrelsen	

Indledning

Udviklingen inden for generativ kunstig intelligens (AI) går stærkt. Efter lanceringen af ChatGPT i oktober 2022 er antallet af generative AI-værktøjer og deres funktionalitet vokset markant. Uanset om jeres organisation har taget stilling til brugen af generativ AI eller ej, så er der en god sandsynlighed for, at flere med-arbejdere i organisationen allerede bruger det – og med god grund.

Generativ AI åbner op for nye måder at løse opgaver på. Hvad end det omhandler effektivisering af tidskrævende manuelle arbejdsopgaver, finde mønstre i store mængder data, komme med nye ideer eller understøtte mere komplekse opgaver, såsom kodning til programmering, er der muligheder.

Faktaboks

Hvad er generativ AI?

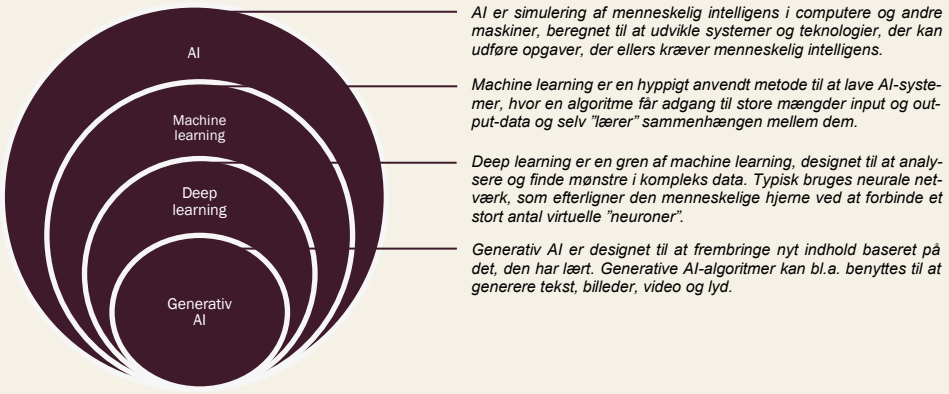
Generativ AI er værktøjer, der udnytter meget store datamængder sammen med maskinlæringsteknikker til at producere indhold, baseret på input fra brugerne kaldet *prompts*. Prompts er de spørgsmål eller kommandoer, brugeren giver til et generativt værktøj for at få et output.

Generativ AI er ikke intelligens i menneskelig forstand. I stedet er det avancerede modeller, der forsøger at forudsige, hvilken tekst, billede, lyd, video eller andet, der med størst sandsynlighed tilfredsstiller den givne prompt. Det er med andre ord (meget) avanceret statistik.

Generativ AI er en delmængde af AI, som bygger på bl.a. teknikkerne *machine learning* og *deep learning*, hvor en algoritme trænes på store datamængder med henblik på at lære den sammenhængen mellem datapunkter. Generativ AI har dog den særlige egenskab, at den er udviklet til at generere nyt indhold på baggrund af de store mængder data, de er trænet på.

Generative AI-værktøjer bygger på såkaldte grundmodeller (på engelsk; ”foundation models”). Grundmodeller er udviklet på en bred mængde af umærkede data (som er tilgængeligt på internettet) og kan tilpasses til mange opgaver. De adskiller sig derved fra mange andre former for AI, som trænes i opgavespecifikke data til at udføre et snævert udvalg af funktioner. Grundmodeller er ofte trænet på både tekst, billede, lyd, video, mv.

Figur 1: Generativ AI er en delmængde af AI



Kilde: McKinsey & Company

Oven på grundmodellerne kan der bygges applikationer, som udgør de generative AI-værktøjer, som brugerne kan interagere med. Mange af disse findes frit tilgængeligt på offentlige hjemmesider eller let tilgængeligt via betalte cloudtjenester. Generative AI-værktøjer indeholder ofte et sæt af regler, som censurerer eller justerer de svar som de giver, fx for at undgå diskriminerende svar, opfordringer til vold eller andet.

Der findes også grundmodeller og generative AI-værktøjer, som er open source, og som derfor kan downloades og afvikles på egne servere. I flere tilfælde er det også muligt at fintune en grundmodel ved at træne videre på den med egne data og finjustere den til en specifik opgave.

Der findes på baggrund af ovenstående en lang række forskellige generative AI-værktøjer, og der kommer hele tiden flere til.

Det er et ledelsesmæssigt ansvar at sikre, at anvendelsen af generativ AI sker på en ansvarlig måde. Til inspiration kan I som organisation gå til opgaven ved

1) først at tage stilling til, hvilken TILGANG til generative AI-værktøjer, der passer til jeres organisations formål,

2) dernæst at lave RETNINGSLINJER tilpas-set jeres valg af generative AI-værktøjer og jeres kontekst med blik for mulighederne, men også med særlig fokus på generelle faldgruber ved generativ AI og de lovgivnings- og forvaltningsmæssige rammer, der er på området, og

3) fastlægge hvordan I som organisation understøtter det i praksis gennem ORGANISATORISKE RAMMER for ibrugtagning af ny teknologi.

Guiden er henvendt til ledere i offentlige myndigheder med henblik på at give inspiration til relevante overvejelser i forbindelse med ibrugtagning af generativ AI. Guiden er en hjælp til at konkretisere anvendelsesformål og -til-

gange, udforme retningslinjer og overveje understøttende organisatoriske tiltag. Det er den enkelte myndigheds ansvar at sikre, at anvendelsen af generativ AI sker ansvarligt.

Guiden vil løbende blive opdateret. Det skyldes både, at den teknologiske udvikling inden for generativ AI går hurtigt, og at der i forbindelse med implementeringen af EU's forordning om kunstig intelligens følger nye rammer og regler, som skal overholdes. Hold jer derfor løbende opdateret på nye versioner af denne guide.

Disse tre trin er illustreret ved billedet nedenfor og gennemgås efterfølgende enkeltvist i denne guide.

1. Tilgang til brugen af generative AI-værktøjer

Der findes mange anvendelsesmuligheder og forskellige typer af generative AI-værktøjer. Det er vigtigt, at I tager stilling til, hvad I ønsker at opnå, hvorvidt generativ AI kan anvendes hertil, og på den baggrund forholder jer til, hvilke krav det stiller til jeres valg af generative AI-værktøjer.

Først er det vigtigt at afgrænse, hvorvidt jeres forretningsbehov dækker over almen anvendelse, specialiseret projektanvendelse eller begge dele.

- **Almen anvendelse af generativ AI.** Med almen anvendelse menes generative AI-værktøjer, som benyttes af alle medarbejdere til flere forskellige formål. Generative AI-værktøjer, som er gratis tilgængelige via offentlige hjemmesider (fx Open AI's ChatGPT), kan i nogle tilfælde være tilstrækkelige til disse formål. Det kan fx være til hurtigt at danne sig et overblik over store mængder offentligt tilgængelig information eller til at generere stykker af programmeringskode ved i almindeligt sprog at beskrive, hvad kodestykket skal kunne. Det kan også være til at generere nye ideer eller komme med udkast. Ved anvendelse af generative AI-værktøjer, som er gratis tilgængelige via en offentlig hjemmeside, er det dog vigtigt at være opmærksom på, at udbyderne ofte har ret til at anvende prompts, metadata og svar i andre sammenhænge, herunder til træning af værktøjerne. Vær derfor opmærksom på beskyttelse af tavshedspligt og personoplysninger (mere herom i afsnit B).

- **Specialiseret projektanvendelse af generativ AI.** Med specialiseret projektanvendelse menes anvendelser, som er afgrænset til specifikke områder og evt. udvalgte medarbejdere. Det kan fx være til at generere indhold eller udarbejde analyser, hvor den generative AI skal have adgang til organisationens viden og tidligere arbejde for at kunne trænes til at imitere en bestemt adfærd eller have domænespecifik viden. Her vil det ofte være nødvendigt at anvende nogle af de ekstra funktionaliteter, som kan tilbydes mod betaling hos flere udbydere af generative AI-værktøjer (fx Microsoft Azure AI) eller drive en model på egne servere. Det kan fx være funktionaliteter til at integrere værktøjerne med organisationens forretningsdata eller integrere med organisationens eksisterende programmer. Et andet behov kan være at købe adgang til de nyeste modeller.

Afhængigt af jeres behov, og de forvaltnings- og lovgivningsmæssige hensyn, der berøres i afsnit B, kan I vurdere, hvilken teknisk løsning, I med fordel kan vælge. Her kan I overveje, hvorvidt jeres behov kan realiseres med en cloudtjeneste eller om I skal drive værktøjet lokalt, samt om I vil vælge en open eller closed source model.

- **Open source vs. closed source.** Nogle generative AI-værktøjer er "open source" (fx Meta's Llama 2 eller Hugging face), mens andre er "closed source" (fx Open AI's ChatGPT eller Google's Gemini), og nogle befinder sig et sted i mellem. Denne distinktion omhandler, hvorvidt AI-værktøjets kode er offentligt tilgængelig, og hvorvidt det er muligt at opnå viden om, hvordan AI-værktøjet er udviklet og fungerer. Gennemsigtigheden med AI-værktøjets opbygning kan være vigtig, fx hvis

det anvendes som beslutningsstøtte til afgørelser. "Open source"-værktøjer giver i forskellig grad mulighed for at teste og modificere det generative AI-værktøj, hvor det med "closed source" ikke vil være muligt.

- **Cloudtjenester vs. lokale servere.** Ønsker man at integrere generative AI-værktøjer med egen data og egne systemer, er det mest udbredte at benytte en cloudtjeneste. Det er dog muligt at downloade "open source"-værktøjer, som kan drives på egne servere. Det kræver dog betydelig teknisk viden og teknisk infrastruktur (servere mv.) at etablere og drive generative AI-værktøjer lokalt. Til gengæld giver et lokalt "open source"-værktøj fuld kontrol over data og mulighed for at tilpasse værktøjet til organisationens specifikke behov.

2. Retningslinjer for brugen af generativ AI-værktøjer

Efter I har taget stilling til hvilke generative AI-værktøjer, der er relevante i jeres organisation og i relation til jeres forretningsbehov, anbefales det, at I udformer retningslinjer, der skal danne rammerne for brugen i jeres organisation.

I den forbindelse er det vigtigt, at I forholder jer til de faldgruber, der kan være ved anvendelse af generativ AI, samt relevante lovgivnings- og forvaltningsmæssige rammer.

Faldgruber ved generative AI-værktøjer

Selvom generativ AI har mange nyttige anvendelsesmuligheder, medfører måden de er opbygget samtidig en række faldgruber, som I med fordel kan forholde jer til. Det gælder særligt:

- **Risiko for faktuelt forkerte svar og hallucination.** Generative AI-værktøjer er bygget til at levere det indhold, som mest sandsynligt imødekommer prompten. Hvis generative AI-værktøjer anvendes til faktuelle opgaver, er det vigtigt at være opmærksom på, at værktøjerne i nogle tilfælde kan komme med forkerte eller opdigtede svar. Det kan bl.a. skyldes, at AI-værktøjets datagrundlag ikke er opdateret eller fyldestgørende på det område, som der spørges til. Mange generative AI-værktøjer er trænet på store dele af det frit-tilgængelige internet, og derfor kan urigtige oplysninger herfra være grundlaget for svar. Derudover kan det være svært at gennemskue, hvorvidt der er tale om opdigtet indhold, da AI-værktøjer er gode til at få det til at fremstå troværdigt. Den skriftlige fremstilling afspejler altså ikke altid, at noget er galt.

- **Risiko for bias.** Generativ AI kan producere svar med bias – det vil sige fordomsfulde eller forudindtagede svar – i relation til fx specifikke køn, etnicitet, politiske holdninger eller andet. Det kan både skyldes bias eller skævheder i det datamateriale, som AI-værktøjet er bygget på, eller at udbyderen af AI-værktøjet kan have indlagt regler, som styrer dens output. Det anbefales derfor, at man er opmærksom på risikoen for bias, når man benytter generative AI-værktøjer – særligt hvis værktøjet er "closed source", hvor der ikke er fuld indsigt i værktøjets opbygning, herunder hvilket datagrundlag, det er trænet og finjusteret på samt hvilke filtre, der strukturerer mønstrene for, hvad værktøjet må foreslå.

- **Informationssikkerhed.** Datalæk er altid en risiko ved brug af onlinetjenester. Det kan både ske som følge af et hackerangreb eller ved et uheld. Det er derfor vigtigt, at der foretages en risikovurdering af den valgte løsning og anvendelsen heraf. Derudover kræver nogle generative AI-værktøjer, at der logges ind. Der bør tages stilling til, hvordan login sker mest sikkert.

Det anbefales, at jeres retningslinjer tager højde for ovenstående faldgruber, fx gennem:

- **Oplysning.** Der kan iværksættes tiltag for at oplyse medarbejdere om de faldgruber, som de skal være opmærksomme på. I den forbindelse kan det overvejes at lave positivlister over de generative AI-værktøjer, der er ledelsesmæssigt godkendt til brug i organisationen. Organisationen kan også give eksempler på anvendelser eller deciderede prompts, som er acceptable eller uacceptable, eller oplyse hvilke typer af data, der må anvendes til generativ AI.

- **Kvalitetskontrol.** Der kan opstilles krav om krydstjek af svar fra generative AI-værktøjer med andre kilder eller på anden vis kvalitetssikre for at undgå faktuelt forkerte svar eller hallucination. Benyttes generativ AI fx til at opsummere tekster eller forklare komplicerede koncepter, forudsætter det, at medarbejderen har en faglig forudsætning for at forholde sig kritisk hertil.

- **Transparens.** Det er ledelsens ansvar, at generative AI-værktøjer anvendes korrekt og til de ønskede formål. Det kan derfor være en god idé at lægge vægt på transparens om anvendelsen af generativ AI. Det kan fx være ved forudgående ledelsesmæssige aftaler inden, at generativ AI anvendes eller ved krav om, at medarbejdere skal tydeliggøre i leverancer, at der er anvendt generativ AI.

- **Separat adgangskode.** Ved anvendelse af generative AI-værktøjer, såvel som andre værktøjer som kræver login, bør der anvendes en separat og sikker adgangskode, som ikke anvendes til andre systemer. På den måde sikrer I, at adgangskoder til flere af jeres systemer ikke lækkes, skulle adgangskoder til et enkelt system blive lækket.

Lovgivnings- og forvaltningsmæssige rammer
Ud over faldgruberne er det vigtigt, at I sikrer jer, at jeres retningslinjer for brug af generativ AI stemmer overens med gældende lovgivnings- og forvaltningsmæssige rammer. Det omfatter både særlige regler på jeres område og derudover særligt databeskyttelsesretten, forvaltningsretten, og immaterialretten.

Det anbefales, at I som minimum forholder jer til følgende:

- **Persondatabeskyttelse.** Myndigheder har en forpligtelse til at sikre håndtering af personoplysninger i henhold til databeskyttelsesretten. Personoplysninger må derfor kun indtastes i generative AI-værktøjer, hvis værktøjet er blevet vurderet og godkendt til brug til personoplysninger. Personoplysninger må derfor ikke indtastes i generative AI-værktøjer, som er gratis tilgængelige via offentlige hjemmesider. Her er det vigtigt at være opmærksom på, at der stadig kan være tale om personhenførbare oplysninger, der også er omfattet af persondatabeskyttelse, selvom man fx fjerner navn fra data, der behandles af det generative AI-værktøj.

- **Tavshedspligt.** Offentligt ansatte har tavshedspligt. Det indebærer, at fortrolige oplysninger, som der opnås adgang til i forbindelse med arbejdet, ikke må videregives til uvedkommende. Fortrolige oplysninger bør kun indtastes i generative AI-værktøjer, såfremt at der er foretaget en risikovurdering af brugen af den valgte løsning. Risikovurderingen bør genbesøges regelmæssigt således den aktuelle anvendelse af AI-værktøjet svarer til de accepterede risici. Fortrolige oplysninger må derfor ikke indtastes i generative AI-værktøjer, som er gratis tilgængelige via offentlige hjemmesider.

- **Gennemsigtighed i forvaltningen.** Derudover skal borgere og virksomheder ifølge forvaltningsloven kunne oplyses om, hvad der ligger til grund for en afgørelse. Da generative AI-værktøjer bygger på komplicerede matematiske modeller, er det på nuværende tidspunkt svært at forklare, hvordan de når frem til deres svar. Det bør derfor overvejes grundigt, hvis svar fra generativ AI anvendes som grundlag for afgørelser over for borgere eller virksomheder. Generativ AI kan dog have mange anvendelser i at understøtte arbejdet med afgørelser, fx ved at frembringe relevante informationer eller undersøge præcedens på området.

- **Retten til ikke at være til genstand for en fuldautomatiseret afgørelse.** Som hovedregel er der ret til ikke at blive underlagt en afgørelse, der alene er baseret på automatisk behandling, som har retsvirkning eller på tilsvarende vis betydeligt påvirker vedkommende. Undtagelser hertil kan dog være tilladt, såfremt vurderingskriterier er entydige eller der kan findes hjemmel i nationalretten eller EU-retten.

- **Immaterialret.** Da generativ AI bygger på eksisterende indhold, kan der genereres resultater, der sammenstykker elementer fra indhold, som kan være beskyttet af ophavsret, varemærker, patenter mv. Området er endnu nyt, og der er derfor usikkerhed, om hvor grænsen går. Der pågår flere retlige stridigheder, som på sigt vil kunne danne præcedens vedrørende immaterialret og generativ AI. Det er dog en god ide ikke at prompte generativ AI til at drage inspiration fra indhold, som er beskyttet af ophavsret, varemærker, patenter mv., da svarene derved vil kunne krænke rettighederne.

3. Organisatoriske rammer for ibrugtagning af ny teknologi

Udover at tage stilling til hvilke formål brugen af generativ AI skal tjene, hvilke AI-værktøjer, der kan anvendes i organisationen og hvilke retningslinjer, der skal gælde for brugen heraf, så er det som altid relevant at tage stilling til, hvordan I som organisation understøtter en ansvarlig anvendelse i praksis gennem organisatoriske rammer.

De rette organisatoriske rammer er essentielt for, at ibrugtagning af generativ AI bliver en succes i jeres organisation. Generelle erfaringer med ibrugtagning af ny teknologi tilsiger, at det er vigtigt fra starten af at tage stilling til bl.a., hvem der skal drive arbejdet, og hvordan de klædes på til det samt prioritering af ressourcer til bl.a. vidensopbygning, udvikling, test og validering. En tommelfingerregel ved indføring af ny teknologi er 80/20-reglen, der siger, at 80 pct. vedrører alt det rundt om – og kun 20 pct. vedrører selve teknologien.

Det anbefales derfor, at I husker at medregne understøttende faktorer fra starten. Det kan fx omhandle praktiske rammer som prioritering af ressourcer ved anvendelsen af generative AI-værktøjer, men også at fordre en god og konstruktiv dialog om erfaringer og overvejelser om brugen af generativ AI i jeres organisation. Jo større gennemsigtighed og åbenhed internt på tværs af organisation, desto bedre muligheder har I som ledelse for at sikre, at jeres rammer og retningslinjer afspejler de behov, erfaringer og opmærksomhedspunkter, der eksisterer i det daglige arbejde hos jeres medarbejdere. Samtidigt muliggør en åben dialog, at I som organisation løbende kan lære af jeres erfaringer og dermed blive endnu skarpere på anvendelsen af generativ AI.

Endelig er det et godt råd løbende at tage jeres valg af generative AI-værktøjer, retningslinjer og organisatoriske rammer op til revurdering, da både teknologien, de lovgivningsmæssige rammer og det kollektive erfaringsgrundlag med brug af generativ AI ændrer sig løbende.