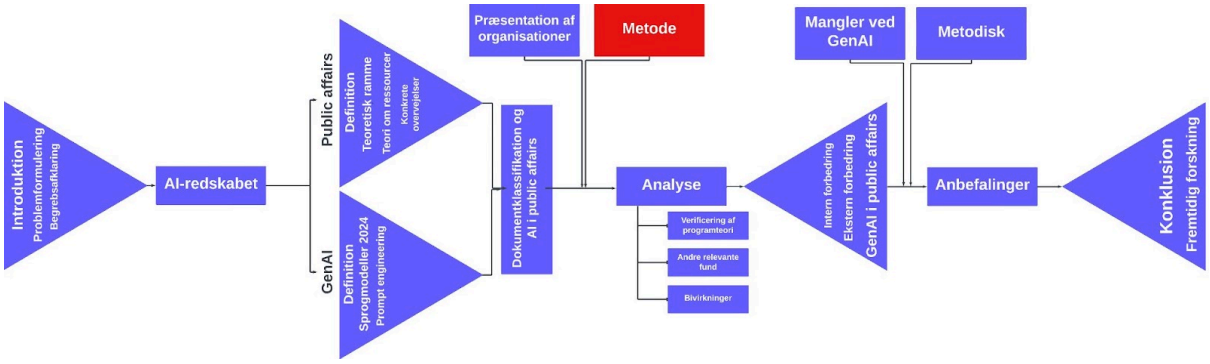
7.0 Metode

Metodeafsnittet kortlægger først kort hvorfor der er benyttet en casetilgang og evaluering som metodisk framework. Dernæst gennemgås valget af procesbaseret virkningsevaluering og programteorien. Det bliver efterfulgt af en kortlægning af min egen rolle i dette speciale samt diskussioner om dataindsamling og -behandling. Metodeafsnittet afsluttes med en diskussion om forskningskriterier.



7.1 Hvorfor en casebaseret tilgang og evaluering

Dette speciale benytter en casebaseret tilgang, som er det specifikke AI-redskab, til at besvare problemformuleringen. Denne tilgang er valgt, fordi sprogmodeller nu er blevet sofistikerede nok til at blive brugt i konkrete redskaber, og fordi der allerede findes mange artikler, der beskriver de potentielle og allerede beviste gevinster ved GenAI (McKinsey, 2023; EY, 2023; BCG, 2023; KPMG, 2023; Chen et al., 2023). Derfor er en specifik case nyttig, da den både kan belyse konkret hvordan GenAI kan skabe værdi og samtidig være så praktisk orienteret, at andre kan bruge fundene til fremtidige undersøgelser af GenAI i public affairs (Andersen et al, 2012).

Valget af evaluering som metodisk tilgang understøtter det overordnede formål med dette speciale. Kun ved at få en evaluering af et AI-redskab, kan man vurdere potentiale samt hvad andre skal være opmærksom på, hvis de ønsker at undersøge andre usecases inden for GenAI i public affairs (Dahler-Larsen, 2018). Det er motiveret af ønsket om en dybdegående forståelse af, hvordan AI-redskabet præsterer i praksis – ikke blot i teorien eller under ideelle omstændigheder og endnu vigtigere: hvor AI-redskabet har mangler.

7.1.1 Procesbaseret virkningsevaluering

Inden for evalueringsfaget findes *procesbaseret virkningsevaluering.* Valget af procesbaseret virkningsevaluering som metodisk tilgang i dette speciale er en måde at undersøge hvordan og hvorfor AI-redskabet potentielt kan forbedre public affairs arbejdet (Dahler-Larsen, 2018).

En række grunde gør virkningsevaluering attraktiv til at besvare problemformuleringen. For det første kan en virkningsevaluering belyse, om AI-redskabet er baseret på en forfejlet idé, eller om det blot kan justeres, hvis AI-redskabet ikke medfører positive outcomes (Dahler-Larsen, 2003). For det andet giver virkningsevaluering mulighed for at give velbegrundede anbefalinger om AI-redskabet på baggrund af fundene fra evalueringen, hvilket vil være afgørende for fremtidige tilgange til design af andre GenAI-løsninger inden for public affairs (Dahler-Larsen, 2018). For det tredje kan virkningsevalueringen belyse *bieffekter*. Bieffekter er relevante effekter, som er utilsigtede eller uforudsete, og derfor ikke blev inkluderet i det oprindelige evalueringsdesign og programteori (ibid). Dette er nyttigt i forbindelse med den eksplorative karakter af dette speciale, da der kan være utilsigtede eller uforudsete effekter, når man evaluerer et helt nyt AI-redskab.

Når man evaluerer, er der objektet eller mekanismen, man forsøger at evaluere, og et outcome, som man vil måle. Men en mekanismes effekt på et outcome er betinget af konteksten, som man opererer i (Pawson & Tilley, 1997). Den samme mekanisme kan producere to forskellige outcomes, fordi det er betinget af konteksten. For at en mekanisme skal lykkes, er det nødvendigt at der er de rigtige betingelser, altså at konteksten stemmer overens med mekanisme og outcome (ibid). I tilfældet af dette speciale handler det om, at effekten og nytten af AI-redskabet er betinget af den organisation, som benytter AI-redskabet. Med dette forstås, at der kan være variation mellem organisationerne på en række centrale parametre som organisationens kapacitet og ønske om at optage ny viden fra Folketinget, hvor centralt behovet er for at være ajour og hvor stort genstandsfeltet er. Konteksten er vigtigt at være opmærksom på, når man skal drage konklusioner på baggrund af virkningsevalueringen.

Et alternativ til den procesbaserede virkningsevaluering er en variansbaseret evaluering. Selvom der er forskellige organisationer, der benytter AI-redskabet, er den procesbaserede tilgang stadig at foretrække, fordi dette speciale er eksplorativt. Derfor er fokuset på processerne ved at benytte AI-redskabet og ikke på variansen af virkningerne ved at benytte

AI-redskabet (Dahler-Larsen, 2018). Det vil sige, at dette speciale ikke fokuserer på at kortlægge, om AI-redskabet virker og i hvilken grad det virker, men hvordan og hvorfor det potentielt virker.

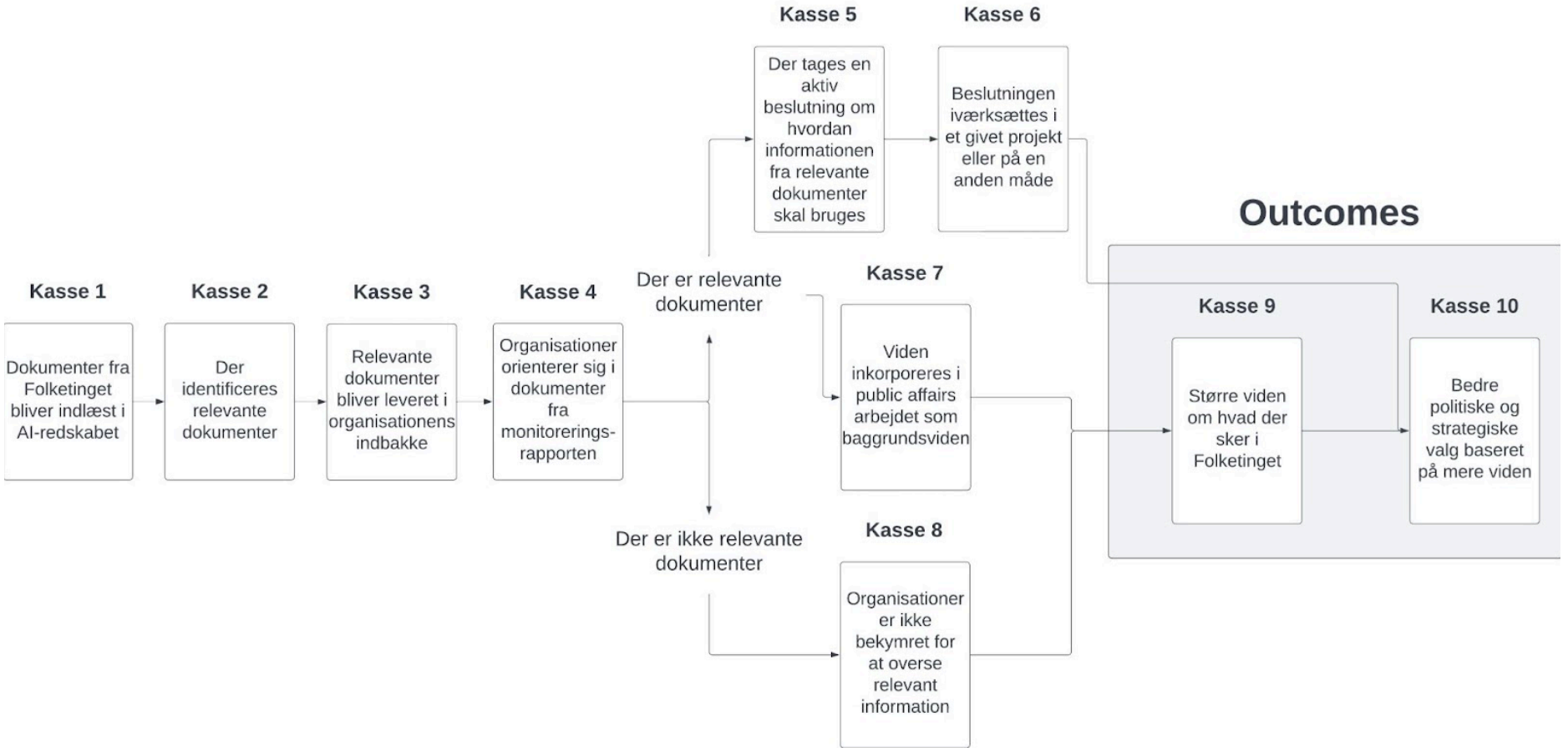
7.2 Programteori

Ved at benytte virkningsevalueringen kan man inddrage og evaluere via en *programteori*. En programteori er: ”begrundede forestillinger om, hvorfor dette projekt skulle have disse resultater for denne gruppe mennesker i denne situation.” (ibid) Programteorien er en måde at strukturere hvordan mekanismerne i AI-redskabet vil medføre ønskede outcomes, hvorfor programteorien kan forstås som et ideal for hvordan et projekt burde fungere (ibid).

Idealet for AI-redskabet er, at AI-redskabet skal gøre det lettere for organisationer at tilgå den store mængde tekstdata, der er tilgængeligt fra Folketinget. Det er et ideal, fordi tekstdataen fra Folketinget ofte indeholder information som er relevant for public affairs organisationer. Det er relevant for public affairs organisationer, fordi det hjælper dem med politisk monitorering ved at holde sig ajour om hvad der bevæger sig i Folketinget. Dermed kan de lettere mappe relevante stakeholders og policy-positioner (Hegelund & Mose, 2013; Esbensen, 2007). Det analoge alternativ til AI-redskabet er, at medarbejderne i public affairs organisationer selv skal gennemgå den store mængde tekstdata fra Folketinget, hvilket er tidskrævende. Derfor er det de færreste, der gør det, og i stedet abonnerer de på relevante folketingsnyhedsbreve, der selv indeholder meget irrelevant information. AI-redskabet vil ideelt kunne frigøre tid til andre mere meningsfulde og værdiskabende opgaver frem for at gennemlæse og manuelt udtrække relevant information fra folketingsdokumenter og sikre, at relevant information fra Folketinget ikke bliver overset.

Programteorien kan ses herunder:

Figur 4: Programteorien



7.2.1 Sproglig forklaring af programteorien

Programteorien kan forklares sprogligt som følgende: AI-redskabet henter og indlæser parlamentariske procesdokumenter (kasse 1) og benytter derefter en sprogmodel til at klassificere hvert enkelt dokument for at afgøre, om det er relevant for en given organisation (kasse 2). Når klassifikationsprocessen er færdig, bliver monitoreringsrapporten sendt ud som en mail til den givne organisation (kasse 3). Herfra er AI-redskabet færdigt med sin proces, og det faktiske arbejde med monitoreringsrapporten begynder i organisationerne. Næste kasse i programteorien er at den givne organisation åbner monitoreringsrapporten og orienterer sig i den (kasse 4). Herfra kan der ske to ting: Enten finder organisationen et eller flere dokumenter, som de mener er relevante for deres arbejde eller der er ikke relevante dokumenter for organisationen. Hvis der er relevante dokumenter, kan der også ske to ting: Enten tager organisationerne en aktiv beslutning om, hvordan informationen fra et dokument skal integreres i deres public affairs arbejde (kasse 5) og derefter iværksætter beslutningen (kasse 6), eller organisationen kan optage viden fra monitoreringsrapporten for at vidensopbygge (kasse 7). Hvis der ikke er relevante dokumenter i monitoreringsrapporten, vil organisationen ikke være bekymret for at have overset relevant information fra Folketinget (kasse 8).

Kasse 1 til 8 medfører to outcomes: For det første medfører kasse 7 og kasse 8, at organisationerne har mere viden om hvad der sker i Folketinget, hvilket er det første outcome af dette AI-redskab (kasse 9). Mere viden om hvad der sker i Folketinget medfører, at organisationerne kan tage beslutninger på et mere oplyst grundlag, hvilket leder til at organisationerne tager bedre politiske og strategiske beslutninger for hvordan public affairs arbejdet skal udføres (kasse 10). Pilen fra kasse 6, der vedrører iværksættelsen af en beslutning baseret på monitoreringsrapporten, fører direkte til kasse 10, som omhandler bedre politiske og strategiske beslutninger. Det skal forstås som at et enkeltstående dokument fra monitoreringsrapporten, der leder til at der bliver iværksat en beslutning, medfører at politiske og strategiske beslutninger bliver bedre, fordi organisationen har fået konkret information fra dokumentet, som de agerer på.

Der er lavet en distinktion mellem kasse 5 og kasse 6, da organisationer kan beslutte at bruge et dokument i deres public affairs arbejde, men der kan være forskellige grunde til at dette aldrig materialiserer sig. I konteksten af AI-redskabet kan det tænkes, at der kommer for

mange dokumenter, hvor organisationerne beslutter at agere på baggrund af et dokument. Hvis der kommer for mange relevante dokumenter, når beslutningen aldrig at blive udført. Hvis det hænger sammen på den måde, skaber AI-redskabet ikke omstændighederne for at tage bedre politiske og strategiske valg.

7.2.2 Forklaring af indholdet af kasserne i programteorien og hvordan de udfyldes

Helt konkret kan en programteori benyttes til at evaluere ved at *verificere* alle kasserne via dataindsamling for at kortlægge en mekanismes effekt (Dahler-Larsen, 2018).

De første 3 kasser i programteorien udfyldes kvalitativt.

7.2.2.1 Kasse 1 til 3: Kvantitative svar

Kasse 1 er ligetil at verificere, da det blot handler om hvorvidt AI-redskabet succesfuldt trækker information fra Folketingets hjemmeside.

Kasse 2 er givetvis den vigtigste af alle kasserne, hvorfor verificeringen af den bliver mest omfattende. Det er i kasse 2, at sprogmodellens klassifikationer bliver verificeret. Denne kasse kan testes kvantitativt ved manuel gennemgang af alle parliamentariske procesdokumenter, hvor antallet af positiver, negativer, falske positiver og falske negativer tælles (Powers, 2021). Derudover laves der to kategorier, hvor der tælles edgecases. *En edgecase for positiver* hvor AI-redskabet har klassificeret et dokument som relevant, men det er uklart om dokumentet faktisk er relevant og en *edgecase for negativer*, hvor AI-redskabet har klassificeret et dokument som ikke-relevant og det er uklart om dokumentet faktisk ikke er relevant. De to kategorier af edgecases er nødvendige, da det ikke altid er helt klart om en sag er relevant eller ikke relevant. For eksempel var der et dokument omhandelnde at der er lukket en læreruddannelse, som der var et spørgsmål til i Børne og Undervisningsudvalget. Uddannelse til velfærdsuddannelser er af interesse for Danske Gymnasier, siden mange velfærdsuddannelser optager elever fra stx, men omvendt ligger læreruddannelsen uden for stx's område. AI-redskabet klassificerede denne sag som relevant for Danske Gymnasier, hvorfor sagen bliver tilknyttet et label som edgecase ved positiv klassificering.

Der tages en stikprøve af AI-redskabets klassifikation fra en tilfældig uge, hvor AI-redskabet havde de endelige prompts. Den tilfældigt valgte periode er er 8-12. april. Som nævnt var det

en iterativ proces at lave prompts i samarbejde med organisationerne, så for at få det mest retvisende billede af AI-redskabets ydeevne, tages stikprøven når promptsene var færdigudviklede.

Det er urealistisk at kortlægge alle AI-redskabets klassifikationer for den fulde implementeringsperiode, da der er omkring 6400 dokumenter og derfor 51.200 individuelle vurderinger af relevans[2].

Kasse 3 er ligetil at verificere. Den handler om, hvorvidt monitoreringsrapporterne er blevet sendt ud.

7.2.2.2 Kasse 4 til 8: Kvalitative svar

Fra kasse 4 skal kasserne verificeres via kvalitativ dataindsamling. Gennem interviews fås indsigt i, om organisationerne læser monitoreringsrapporten (kasse 4), om de har opfanget dokumenter, som de mener, at de skal agere på (kasse 5), om de faktisk har iværksat beslutningen om at agere (kasse 6), om de bruger dokumenterne fra monitoreringsrapporten til at få mere viden (kasse 7) og om de ikke er bekymret for at overse relevante dokumenter, hvis der ikke er nogle relevante dokumenter i monitoreringsrapporten (kasse 8).

7.2.2.3 Kasse 9: Mere viden om hvad der sker i Folketinget

Kasse 9 er første outcome. Det er i tråd med afsnit 3.1. i den konceptuelle kontekst, hvor det fremhæves, at for at kunne bedrive god public affairs, bliver man nødt til at vide, hvad der sker i Folketinget (Hedelund & Mose, 2013; Esbensen, 2007). Dette undersøges kvalitativt, hvor der gennem interviews spørges ind til om organisationerne ved mere om hvad der sker i Folketinget inden for deres givne område på baggrund af den information, som organisationerne får gennem monitoreringsrapporterne. Det er vigtigt at isolere, at det er specifikt AI-redskabet der har tilvejebragt den øgede viden om, hvad der sker i Folketinget, fordi hvis al viden fra AI-redskabet også kan fås fra deres nuværende arbejdsgange med at modtage nyhedsbrevene fra Folketinget, er det ikke AI-redskabet, der har medført mere viden om hvad der sker i Folketinget. Det skal altså være viden, som de ellers ikke havde modtaget eller været opmærksom på.

[2] Matematikken er: Omkring 160 dokumenter per dag * 8 prompts (4 organisationer, hvor 2 af dem har 3 prompts) * 5 dage om ugen * 4 uger på en måned * cirka 2 måneder live =
51.200 vurderinger for 6400 dokumenter.

7.2.2.4 Kasse 10: Bedre politiske og strategiske beslutninger baseret på mere viden

Kasse 10 er det andet outcome. Der er ikke meget ved at vide mere om hvad der sker i Folketinget, hvis det ikke medfører at organisationerne bliver bedre til at udøve deres public affairs arbejde. Ellers er AI-redskabet blot et gimmick og ikke et nyttigt redskab. Helt grundlæggende medfører bedre politiske og strategiske beslutninger, at public affairs organisationer kan positionere sig bedre i den kompetitive pluralistiske public affairs tradition i Danmark, så de kan maksimere deres indflydelse (Hedelund & Mose, 2013; Esbensen, 2013). Dette undersøges kvalitativt, hvor der spørges ind til om organisationerne, på baggrund af den viden de har fået gennem AI-redskabet, har været i stand til at tage bedre beslutninger både i forhold til politik og i forhold til strategi. Bedre beslutninger i forhold til politik skal forstås som, at organisationen har indtaget eller justeret en holdning efter at have fået mere viden om, hvad der sker i Folketinget. Bedre beslutninger i forhold til strategi skal forstås som at organisationen har taget nogle strategiske valg, der har til mål at fremme deres public affairs arbejde. Det kunne for eksempel være at organisationen har lagt en strategi om at samarbejde med en bestemt politiker på baggrund af et udvalgsspørgsmål, eller at organisationen har taget et strategisk valg om at forfølge en dagsorden baseret på et §-20-spørgsmål.

7.2.3 Implementeringsfejl og teorifejl

Ved at benytte en programteori kan man med større sikkerhed fastslå om AI-redskabet lider af *implementeringsfejl*, *teorifejl* eller begge dele, hvilket er essentielt, da det vil være arrogant at tro, at AI-redskabet ikke rummer begge dele. Implementeringsfejl indebærer, at AI-redskabet er blevet implementeret mangelfuldt, så den ikke svarer til det, som programteorien foreskriver at AI-redskabet skal (Dahler-Larsen, 2003). Det kunne eksempelvis være, hvis monitoreringsrapporten ikke bliver sendt ud eller monitoreringsrapporten er for uoverskuelig, så den ikke bliver læst. Teorifejl er hvis AI-redskabet følger programteorien, men at outcome alligevel er udeblevet (ibid). Det kunne være hvis AI-redskabet gør hvad den skal som følge af den opstillede programteori, men den stadig ikke leder til mere viden om hvad der sker i Folketinget eller bedre strategiske eller politiske valg i organisationerne. Hvis AI-redskabet er implementeret korrekt og AI-redskabet har medført positive outcomes som foreskrevet i programteorien, er programteorien styrket. Hvis AI-redskabet er implementeret korrekt, men positive outcomes er udeblevet, kan man konkludere teorifejl. Hvis AI-redskabet er

implementeret mangelfuldt, men resultatet alligevel er indtruffet, er der andre forhold end AI-redskabet, der har medført de positive outcomes, hvorfor man ikke har nok data til at ændre i sin tiltro til programteorien. Slutteligt, hvis AI-redskabet er implementeret mangelfuldt og positive outcomes er udeblevet er der implementeringsfejl og måske teorifejl (ibid). Det er det sidste scenarie, som forekommer hyppigst i virkningsevalueringer, selvom det til tider kan være svært at præcist definere, hvorvidt noget er en teorifejl eller en implementeringsfejl (ibid).

7.2.4 Tidsbegrænsning i programteori

Der er en naturlig tidsbegrænsning af programteorien givet specialeperioden, som betyder at AI-redskabet blot når at være implementeret i omkring 2 måneder (Dahler-Larsen, 2018). Det betyder for det første, at de folketingsdokumenter, som de relevante dokumenter omhandler, potentielt ikke når at blive færdigbehandlet inden for perioden af specialet, hvorfor det fulde potentiale for AI-redskabet ikke når at blive forløst. Det kan tænkes, at der kan blive stillet et udvalgsspørgsmål, men der når ikke at komme et svar på spørgsmålet inden for tidsbegrænsningen, hvor svaret er af relevans for organisationerne. For det andet betyder det også, at der muligvis ikke når at komme nok relevante dokumenter, som AI-redskabet kan klassificere, til at AI-redskabet viser sin nytte, hvorfor der potentielt vil være utilstrækkelig data til at evaluere om der er en effekt af at benytte AI-redskabet. For det tredje kan tidsbegrænsningen have betydning for brugernes tilvænning til AI-redskabet, hvor der skal opbygges tillid til at AI-redskabet opfanger det, som den skal opfange.

7.2.5 Begrænsning af outcome i programteori

Når man udformer en programteori, skal man foretage et valg om hvor langt ud i outcome man vil gå (Dahler-Larsen, 2018). Afgrænsningen i outcome er begrænset til, at organisationerne tager bedre politiske og strategiske valg. Det er gjort af to grunde: For det første er det notorisk svært at måle effekter af public affairs, hvorfor det ville være en udfordring at måle på hvorvidt organisationernes public affairs arbejde har virket (Ferraro, 2000). Selvom effekten af public affairs arbejdet kunne måles, ville det være udfordrende at isolere at public affairs arbejdet er blevet bedre *på grund af* AI-redskabet. Det vil være en langt større undersøgelse end dette speciale tillader. For det andet, givet formålet for dette speciale, prioriteres det at kunne konkludere noget konkret med stor sikkerhed frem for at

skulle lave en række antagelser. Ved at kunne konkludere noget konkret, er det lettere for andre at bygge oven på og blive klogere på præcis hvordan AI-redskabet fungerer og hvordan det kan understøtte og forbedre organisationers arbejde med public affairs, hvilket også ligger i tråd med problemformuleringen i dette speciale.

7.3 Min egen rolle

Siden AI-redskabet er designet og udviklet af mig selv, rejses der uundgåeligt spørgsmålet om objektivitet og udvikler-bias, idet jeg står over for opgaven at evaluere et redskab, jeg selv har lavet (Chattopadhyay, 2022). Derudover kan min tilknytning til de fire organisationer også medføre biased dataindsamling og konklusioner. På trods af disse potentielle bekymringer, anses objektivitetsudfordringen ikke at være en begrænsende faktor for at gennemføre specialet, men er det noget som jeg i høj grad skal have in mente, når der skal drages konklusioner.

Objektivitetsudfordringen vurderes ikke til at være en begrænsende faktor for gennemførelsen af specialet af tre grunde: For det første er formålet med specialet at være en af de første undersøgelser om hvordan GenAI kan forbedre public affairs, som synes at vægte højere end objektivitetsudfordringerne, der kan mitigeres med robuste metodiske valg og gennemsigtighed (ibid). For det andet giver min involvering i projektet en større forståelse for både hvordan AI-redskabet fungerer og hvordan organisationerne arbejder. Derfor, selvom min involvering uden tvivl tilføjer et lag er subjektivitet, vurderes det, at det til gengæld medfører en dybere forståelse og perspektiver, som kan anvendes til at nuancere og forbedre evalueringen af AI-redskabet. For det tredje kan jeg gennem en tilknytning til den akademiske litteratur om evaluering og gennemsigtighed i forhold til de metodiske valg opretholde en nødvendig distancering og justere konklusionerne som følge af objektivitetsudfordringen.

Slutteligt er det også relevant at have in mente når man skal drage konklusioner på evalueringen af AI-redskabet, at jeg fik ideen til at bygge et værktøj som AI-redskabet mens jeg arbejdede i Danske Gymnasier, da der var et relevant dokument som jeg overså, fordi det var under Finansudvalget og ikke Børne- og Undervisningsudvalget. Teknologien var dog ikke moden til at bygge AI-redskabet dengang. Dette faktum er vigtigt at have in mente, fordi selvom jeg har forsøgt at designe AI-redskabet til universel brug på tværs af

organisationsstyper, har jeg taget valg baseret på min viden om hvordan arbejdsgangen er i Danske Gymnasier. Valgene er både bevidste, som at formen skal være en daglig orienteringsmail, eller ubevidste. Dette faktum skal inddrages, når der drages konklusioner.

7.4 Dataindsamling

Dataindsamlingen til dette speciale hviler primært på interviews med organisationerne som benytter AI-redskabet, da det er prioriteret at kunne indrette indsamlingen af data baseret på de forskellige organisationers behov og kontekster samt at interview som dataindsamlingsmetode giver mulighed for at beskrive og forstå hvordan brugere oplever at benytte AI-redskabet (Harrits et al, 2012). Der er sikret engagement fra fire organisationer, som alle har forpligtet sig til at deltage i tre runder af interviews hver over specialeperioden.

7.4.1 Første runde af interviews

Første runde af interviews var en indledende dataindsamling, hvor der blev spurgt ind til organisationernes førstehåndsindtryk ved at benytte AI-redskabet og om AI-redskabet havde bidraget til at forbedre deres public affairs arbejde via mere viden. Derudover blev der spurgt ind til om dokumenterne, som AI-redskabet klassificerede som relevante, faktisk var relevante, med henblik på at justere promptet via den iterative proces som beskrevet i afsnit 4.3. Selvom det på forhånd var undersøgt og kortlagt hvad der var relevant for de enkelte organisationer ved at snakke med dem og læse om deres arbejde, kunne deres ønsker til hvordan AI-redskabet håndterede falske positiver have ændret sig efter de havde benyttet AI-redskabet og læst monitoreringsrapporten. Det kunne for eksempel være, at de ønskede at skrue op eller ned for antallet af falske positiver.

7.4.2 Anden runde af interviews

Anden runde af interviews var en stringent evaluering af programteorien. Her indsamles data om AI-redskabets ydeevne og om hvorvidt det har været til nytte for organisationerne. Konkret var der et spørgsmål tilknyttet til hver kasse af programteorien, som organisationerne skulle svare på.

7.4.3 Tredje runde af interviews

Tredje og sidste runde af interviews havde til mål at sikre, at fundene fra anden runde af interviews blev dobbelttjekket. Det blev prioriteret at dobbelttjekke, fordi organisationernes holdning til om hvorvidt AI-redskabet havde været nyttigt kunne have ændret sig. Denne stringente tilgang øger validiteten af de potentielle fund og sikrer at ikke blot en eller to monitoreringsrapporter danner organisationernes holdning, men at potentielle positive evalueringer af AI-redskabet er funderet i længerevarende brug. Med andre ord skal det sikres, at det ikke blot er tilfældigheder, der udgør organisationernes svar til evaluering.

Et andet mål med tredje runde af interviews var at undersøge om der var andre muligheder for at benytte GenAI i public affairs. Problemformuleringen bruger som sagt blot AI-redskabet som case til at undersøge spørgsmålet. Efter at have brugt AI-redskabet, havde organisationerne en vis forståelse for, hvordan GenAI fungerer, hvilket tillader dem at danne mere kvalificerede holdninger til hvordan GenAI kan bruges til at forbedre public affairs.

7.4.4 Mætningspunkt

Der blev nået et mætningspunkt for dataindsamling hurtigere end forventet, da tre ud af fire organisationer ikke læste monitoreringsrapporten fast, hvorfor det var uhensigtsmæssigt og nytteløst at interviewe dem yderligere (Harrits et al, 2012). Helt konkret endte det med at kun Danske Gymnasier nåede at gennemføre alle tre interviews, da de som eneste organisation fast læser monitoreringsrapporten. Der blev gennemført ét interview med SJ&K med opfølgende spørgsmål i et kort efterfølgende telefonopkald. I det ene interview blev SJ&K også spurgt ind til hvordan de så fremtidig brug af GenAI inden for public affairs. Der blev gennemført to interviews med Zero Carbon Shipping og to interviews med Dansk Standard. Det betød, at interview nummer to hos Zero Carbon Shipping og Dansk Standard blev brugt til at udforske hvordan fremtidig brug af GenAI kan forbedre public affairs.

7.4.5 Forud for de formelle interviews

Forud for de formelle dataindsamlingsinterviews blev der afholdt et møde med hver organisation for at præsentere AI-redskabet og spørge om de ville indgå et samarbejde med henblik på dataindsamling til specialet. Disse møder var overvejende introducerende og havde til formål at afklare, hvad hver organisation specifikt søgte at få en monitorering af via

AI-redskabet. Der blev hverken taget noter med henblik på specialet eller optaget, så indholdet fra møderne var udelukkende for den efterfølgende udarbejdelse af prompt og samarbejdsaftale. Intet fra de indledende møder er inkluderet i specialet, og hvis der blev sagt noget interessant i de indledende møder, er det blevet verificeret via interviews, så alt data findes i transskriptionerne.

7.4.6 Inddragelse af data scientist som ekspertkilde

Udover interviews med brugerne af AI-redskabet er Simon Moe Sørensen, en AI-ingeniør og ekspert fra konsulentfirmaet 2021.ai, også blevet inddraget som supplerende ekspertviden om hvordan AI-redskabet er designet (Harrits et al, 2012). Han blev inddraget efter AI-redskabet blev implementeret, og har derfor ikke indflydelse på udformningen og designet af AI-redskabet. Interviewet med Sørensen har til formål til at kvalificere udsagn om AI-redskabets design og ydeevne yderligere, hvilket styrker troværdigheden af dette speciale og giver rygdækning til at argumentere at enten kan AI-redskabet blive bedre eller at udnyttelsen for sprogmodellen i AI-redskabet er på det maksimale. Denne viden om AI-redskabets ydeevne har betydning for konklusionerne om fremtidig design af GenAI-redskaber. Hvis AI-redskabet ikke kan blive bedre, kan det konkluderes, at man skal vente på at teknologien bliver bedre, hvis den bliver bedre. Hvis AI-redskabet godt kan blive bedre, åbner det for muligheder for at andre kan undersøge og udforske GenAI i public affairs i fremtidig forskning.

7.5 Databehandling og analysestrategi for interviews

I processen med at bearbejde og analysere data fra interviewene er der blevet fulgt en iterativ tilgang, som har muliggjort løbende vidensgenerering gennem hele specialeperioden, hvilket er nødvendigt i et eksplorativt studie (ibid). Selvom det var en mulighed at udarbejde tre faste interviewguides fra starten og fulgt disse slavisk, blev der valgt en mere fleksibel strategi for ikke at overse nuancer fra interviews. Ved at bevare de samme overordnede tematikker gennem hver interviewrunde, men tillade justeringer baseret på ny viden og konteksten for hver organisation, kunne der udtrækkes maksimal viden ud af hvert interview, mens sammenligneligheden på tværs af organisationer stadig var mulig. Dermed var det en kumulativ tilgang til indsamling af interviewdata (Elklit & Jensen, 2012). Interviewguides og transskriberinger kan ses i bilag 6 til 14 og 18 til 26 henholdsvis.

For at strukturere og analysere interviewdataen blev Nvivo anvendt som en systematisk støtte i kodningsprocessen af datamaterialet. En start- og slutkodeliste dannede rammen om denne kodningsproces, mens overvejelser omkring intrakode-reliabilitet sikrede konsistens og pålidelighed i kodingsarbejdet med henblik på at øge gennemsigtigheden ved de metodiske valg og hvordan interviewdata kan bruges til at drage konklusioner (Jakobsen, 2012). Start- og slutkodelisten findes i bilag 15 og 16 og intrakode-reliabiliteten er på 71% i Cohens kappa koefficient, hvilket er et mål, som tager højde for sandsynligheden i sammenfaldet af kodning. 71% er en solid intrakode-reliabilitet uden at være fremragende, men acceptabelt når der ikke var faste interviewguides og iterativ vidensgenerering (Nvivo, 2024).

7.6 Vurdering af udvalgte forskningskriterier

Ved en diskussion af udvalgte forskningskriterier, forsøges det at imødekomme kritikpunkter vedrørende dette speciales forskningsdesign.

7.6.1 Intern validitet

Intern validitet refererer til spørgsmålet om, hvorvidt resultaterne, der er fundet via interviews, faktisk afspejler det, som der er forsøgt at blive undersøgt (Andersen, 2012). I konteksten af dette speciale handler det om, at hvis det bliver vist, at AI-redskabet klassificerer dokumenter fra Folketinget, der er relevante for organisationerne, skal det være dokumenter, som organisationerne ellers ikke havde været opmærksomme på. Ellers vil det være en anden effekt end AI-redskabet, der øger vidensniveauet i organisationerne.

7.6.2 Ekstern validitet

Den eksterne validitet ved fundene i dette speciale omhandler, hvorvidt man kan bruge AI-redskabet som case til at generalisere svaret ud på problemformuleringen; altså om man kan svare på, hvilket potentiale der er at benytte GenAI i public affairs (ibid). Ved at snævre strukturen af dette speciale ned til en case, kan man selvsagt ikke give et fuldstændigt svar på det fulde potentiale ved at benytte GenAI i public affairs, da dette emne i sin helhed er for stort til at blive undersøgt i ét speciale. Men ved at undersøge en delkomponent af det, kan det belyses hvor god GenAI er til dokumentklassifikation, som kan være en vigtig del af public affairs (Esbensen, 2007), som fremtidig forskning kan undersøge, der kan bruges til at forbedre public affairs.

7.6.3 Reliabilitet

Et lille, men vigtigt punkt for dette speciale er reliabilitet, som betyder at ved gentagne målinger får man samme resultat (Andersen, 2012). Det er vigtigt at gøre klart, at AI-redskabet *er* konsistent i klassifikationen af dokumenter. Siden sprogmodeller er autoregressive og stokastiske, og derfor baserer sig på sandsynlighedsregning jf. afsnit 4.2, kan det ikke udelukkes, at der kan være forskellige klassifikationer af det samme dokument, hvis man tester på det samme datasæt af dokumenter flere gange. Efter at have gjort det, fandt jeg at der *ikke* er forskel på to gennemgange af det samme datasæt for forskellige organisationer. Datasættet og reliabilitetstesten findes i bilag 17.

7.6.4 Replikerbarhed

Det sidste forskningskriterie er at andre skal kunne genskabe samme resultater ved at benytte samme metode, hvilket er ekstra relevant for dette speciale på grund af udvikler-bias (ibid). Replikerbarheden er øget i dette speciale ved at klart beskrive AI-redskabets arkitektur, vedlægge de benyttede prompt, vedlægge de fulde transskriptioner, vedlægge interviewguides og vedlægge kodelisten.

Der er dog to ting der taler imod at andre kan replikere dette studie: For det første er organisationerne udvalgt på baggrund af mit netværk, som selvsagt ikke kan replikeres af andre. Dette anses dog ikke som et stort problem, da andre organisationer, der beskæftiger sig med public affairs, vil kunne lave samme evaluering ved at svare på de samme spørgsmål. For det andet, og en potentielt større barriere for replikerbarhed, er at det generelt er en udfordring at have høj replikabilitet i kvalitative studier. Det er dog en afvejning, da kvalitative studier ofte medfører høj intern validitet, hvilket er mere ønskværdigt i et eksplorativt studie som dette (ibid).