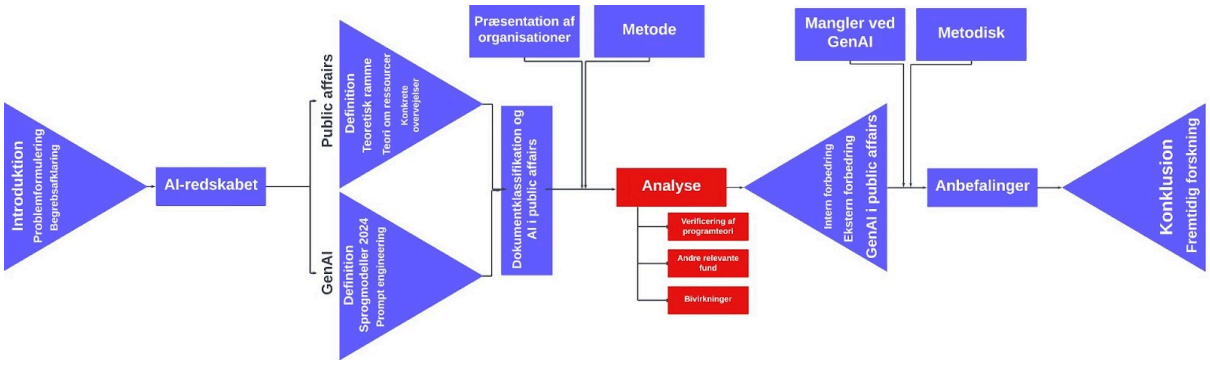
8.0 Analyse

Analysen er struktureret i tre dele: Først gennemgås programteorien kasse for kasse for at få en stringent evaluering af AI-redskabet som leder hen til vurderinger om teorifejl og implementeringsfejl. Dernæst fremlægges relevante observationer om AI-redskabet, som kom til udtryk gennem dataindsamlingen, men ikke er inkluderet i programteorien eller falder under bivirkninger. Relevante observationer er observationer, som fremtidige forskere og praktikere kan have in mente, hvis de ønsker at undersøge potentialet for GenAI i public affairs. Slutteligt fremlægges bivirkninger som blev identificeret som følge af dataindsamlingen (Dahler-Larsen, 2018).



8.1 Evaluering af programteorien

For overskuelighed findes en nummerering af bilag, tilhørende interview og en forkortelse af referencen her. Den bruges til at lave referencer i teksten til interviews:

Tabel 2: Bilagsoversigt for interviews

Bilag	Organisation og runde interview	Forkortelse der bruges som henvisning
Bilag 18	Danske Gymnasier 1	DG1
Bilag 19	Danske Gymnasier 2	DG2
Bilag 20	Danske Gymnasier 3	DG3
Bilag 21	Dansk Standard 1	DS1
Bilag 22	Dansk Standard 2	DS2
Bilag 23	SJ&K 1	SJ&K1
Bilag 24	SJ&K 1,5 (et opfølgende spørgsmål)	SJ&1,5

Bilag 25	Zero Carbon Shipping 1	ZCS1
Bilag 26	Zero Carbon Shipping 2	ZCS2

8.1.1 Kasse 1: Dokumenter fra Folketinget bliver indlæst i AI-redskabet

Den her kasse er ikke interessant, men helt grundlæggende for AI-redskabets virke.

Der har kun været to nedbrud: ét hvor der var nedbrud fra Folketingets Åbne Data og ét hvor Finansministeriets pressemeddelsesdatabase medførte fejl. Det første nedbrud betød at der ikke kom mails ud i 2 dage (15 og 16. januar), hvorfor det kun var SJ&K, der oplevede nedbruddet. Det andet var en aften, hvor der blev testet et prompt, hvor Finansministeriets pressemeddelelsesservice medførte fejl. Siden det skete om aftenen, havde det ingen betydning, da ingen organisationer oplevede nedbruddet.

Derfor kan denne kasse bekræftes som **verificeret**, selvom der var et enkelt nedbrud.

8.1.2 Kasse 2: Der identificeres relevante dokumenter

Denne kasse er central og yderst relevant, da den viser sprogmodellens ydeevne til at klassificere dokumenter korrekt. Derfor er verificeringen af denne kasse den mest omfattende. Verificeringen er delt op i følgende afsnit: Først præsenteres den manuelle proces til at kategorisere og give labels til dokumenter. Derefter introduceres måleenhederne *præcision* og *sensitivitet, som* leder hen til et scoreboard af AI-redskabets klassifikationer. Derfra analyseres scoren fra scoreboardet via overvejelser om genstandsfeltets størrelse og organisationernes forskellige kapaciteter.

8.1.2.1 Yderligere introduktion til processen om at manuelt tildele labels til dokumenter

Der gennemgås en stikprøve af dokumenter manuelt, og denne stikprøve sammenlignes med AI-redskabets klassifikationer. På baggrund af den manuelle klassifikation gives der labels til AI-redskabets klassifikationer på følgende måde, som først introduceret i afsnit 7.2.2.1:

- ➢ Sande positiver: Relevante dokumenter, der korrekt er klassificeret som relevante.
- ➢ Sande negativer: Ikke-relevante dokumenter, der korrekt er klassificeret som ikke-relevante.

- ➢ Falske positiver: Ikke-relevante dokumenter, der fejlagtigt er klassificeret som relevante.
- ➢ Falske negativer: Relevante dokumenter, der fejlagtigt er klassificeret som ikke-relevante.
- ➢ Edgecases ved positiv klassifikation: Det er uklart, om dokumentet er relevant, hvor AI-redskabet har klassificeret det som relevant.
- ➢ Edgecases ved negativ klassifikation: Det er uklart, om dokumentet er relevant, hvor AI-redskabet har klassificeret det som ikke-relevant.

Denne metode til at give labels til klassifikationerne med edgecases er fordelagtig fordi den giver mulighed for at fange nuancer, da det til tider er uklart, om et dokument er relevant jf. eksemplet om den lukkede læreruddannelse fra afsnit 7.2.2.1. Derfor er det nyttigt, at det manuelle klassifikationssystem tillader en fleksibilitet i vurderingen af relevans.

En alternativ tilgang kunne være at gentage en optælling uden edgecases flere gange for at måle en form for intra-koderelabilitet. Denne metode ville dog også medføre udfordringer, som hvordan forskellene mellem optællingerne skal håndteres, og problemet med at lave en skarp opdeling af noget, der i sin essens er nuanceret, vil stadig ikke være løst.

8.1.2.2 Introduktion til præcision og sensitivitet

Ovenstående tilgang til klassifikation giver mulighed for at beregne præcision og sensitivitet, som er standardberegninger i klassificeringsproblemer (Powers, 2011). Præcision beregnes som antallet af sande positive klassifikationer divideret med summen af sande positive vurderinger og falske positiver. Dette måler, hvor stor en andel af de dokumenter, AI-redskabet har klassificeret som relevante, der faktisk er relevante.

$$Præcision = \frac{Antal\ sande\ positive\ klassifikationer}{Antal\ sande\ positive\ klassifikationer + antal\ falske\ positive\ klassifikationer}$$

Sensitivitet beregnes som antallet af sande positive klassifikationer divideret med summen af sande positive klassifikationer og falske negative. Dette måler, hvor stor en andel af de faktisk relevante dokumenter, der korrekt er klassificeret som relevante af AI-redskabet.

$$Sensitivitet = \frac{Antal\ sande\ positive\ klassifikationer}{Antal\ sande\ positive\ klassifikationer + antal\ falske\ negative\ klassifikationer}$$

Ofte opstår der en afvejning mellem præcision og sensitivitet (ibid). Dette kan sammenlignes med at justere maskernes størrelse på et fiskenet. Hvis maskerne er store, minimerer man risikoen for at fange skrald, men risikerer at nogle fisk slipper igennem, hvilket illustrerer høj

præcision. Hvis maskerne derimod er meget små, sikrer man at fange alle fisk, men det kan også betyde, at man fanger skrald sammen med fiskene, altså en høj sensitivitet. Dette medfører en afvejning om, hvor mange fisk – eller i dette tilfælde relevante politiske dokumenter – man er villig til at lade undslippe. Organisationerne skal med andre ord gøre op med sig selv, om de mener, at falske negativer er uacceptable, eller om det er acceptabelt, at der er risiko for falske negativer, hvis det kan reducere antallet af støj, som er falske positiver.

Siden det er en afvejning mellem præcision og sensitivitet, hvor det er op til de enkelte organisationer at vurdere, hvor de vil lægge snittet, sættes der ikke foruddefinerede benchmarks op. Et benchmark kunne for eksempel være at måle hvor mange falske positiver der er i nyhedsbrevene fra Folketinget, siden ikke alt fra et udvalg er relevant for en given organisation, men så skulle man læse alle andre nyhedsbreve fra Folketinget for at kortlægge falske negativer. Det ville skabe sammenlignelighed mellem standardtilgangen med nyhedsbreve og AI-redskabet, men det er et arbitrært sat benchmark, som ikke vil bidrage med substantiel værdi i evalueringen af AI-redskabet, fordi det i sidste ende er op til de enkelte organisationer at vurdere hvad der er relevant og om monitoreringsrapporten lever op til deres behov. Derfor skal resultaterne af den manuelle optælling i kasse 2 ses i sammenhæng med organisationernes svar i kasse 4 til 10.

8.1.2.3 Resultat af manuel klassificering

Fra stikprøven 8. til 12. maj findes følgende resultater:

Tabel 3: Stikprøveresultater af manuel klassificering

	Tech-virksomhed	Social-organisation	Fødevare-virksomhed	Zero Carbon Shipping*	Dansk Standard Byggeri	Dansk Standard Energi	Dansk Standard Standarder**	Danske Gymnasier
Sande positiver	21	12	1	12	3	10	1	5
Sande negativer	764	753	785	173	783	774	801	780
Falske positiver	7	26	10	7	17	11	1	2
Falske negativer	2	1	0	0	0	2	0	2
Edgecases ved positiv klassificering	7	7	2	4	0	4	0	6
Edecases ved negativ klassificering	2	4	5	0	0	2	0	8

N	*803*		*803*		*803*		*196*		*803*		*803*		*803*		*803*	

** AI-redskabet filtrerede kun efter EU-dokumenter, der blev processeret af Folketinget, hvorfor N er lavere hos Zero Carbon Shipping og der er en anden periode (18-22. marts) for den manuelle evaluering, da samarbejdet sluttede med Zero Carbon Shipping før promptet var færdigudviklet hos de andre organisationer.*

*** Dansk Standard ønskede en monitorering af hvor ordet "standard" indgik, men siden der kun er 1 sand positiv og 1 falsk positiv, bearbejdes denne del af deres monitorering ikke yderligere.*

Selvom man kan beregne præcision og sensitivitet baseret på tallene ovenfor, er det netop edgecases, der kan have betydning for hvorvidt AI-redskabets ydeevne er tilfredsstillende.

Man kan tolke edgecases negativt og positivt. Negativ tolkning betyder at edgecases ved positiv klassificering bliver lagt i falske positiver og edgecases ved negativ klassificering bliver lagt i falske negativer. Positiv tolkning betyder at edgecase ved positiv klassificering bliver lagt i sande positiver og edgecases ved negativ klassificering bliver lagt i sande negativer. Med andre ord får man tallene for præcision og sensitivitet til enten at være de bedst mulige eller de værst mulige.

Tabel 4: Oversigt over præcision og sensitivitet ved både positiv og negativ tolkning

	Techvirksomhed		Social-organisation		Fødevare-virksomhed		Zero Carbon Shipping		Dansk Standard Byggeri		Dansk Standard Energi		Dansk Standard Standarder		Danske Gymnasier	
Tolkning	Positiv	Negativ	Positiv	Negativ	Positiv	Negativ	Positiv	Negativ	Positiv	Negativ	Positiv	Negativ	Positiv	Negativ	Positiv	Negativ
Præcision	80%	60%	42%	20%	23%	8%	70%	52%	15%	15%	56%	40%	50%	50%	85%	38%
Sensitivitet	93%	84%	95%	71%	100%	17%	100%	1000%	100%	100%	88%	71%	100%	100%	85%	36%

Det sande tal for hhv. præcision og sensitivitet er et sted mellem de to tolkninger, hvorfor der er taget et gennemsnittet af de to beregninger af præcision og sensitivitet som en proxy for det sande tal. Igen skal det huskes på at det er en afvejning mellem præcision og sensitivitet, hvor det er op til de enkelte organisationer at vurdere, hvor de ønsker at snittet i afvejningen skal være. Efter at have taget gennemsnitsberegningen, kan man lave et scoreboard, som et simpelt men brugbart mål, hvor man summerer præcision og sensitiviteten ved gennemsnittetsberegningen, hvor det højeste tal er bedst.

Tabel 5: Gennemsnitsberegningen mellem positiv og negativt tolkning og scoreboard

	Tech-virksomhed	Social-organisation	Fødevare-virksomhed	Zero Carbon Shipping	Dansk Standard Byggeri	Dansk Standard Energi	Dansk Standard Standarder	Danske Gymnasier
Præcision	70%	34%	15%	61%	15%	48%	50%	62%
Sensitivitet	89%	83%	58%	100%	100%	79%	100%	60%

Sum af præcision og sensitivitet	159	117	74	122	115	127	150	161

8.1.2.4 Scoreboard og hvilket prompt der klarer sig bedst

Det ses af scoreboardet at Zero Carbon Shipping og Techvirksomheden er de organisationer, hvor det tilhørende prompt performer bedst i AI-redskabet. Generelt kan man se, at AI-redskabet ikke er særligt god til at klassificere relevante dokumenter fra Folketinget, og at der er for mange falske positiver, som afspejles i den generelt lave præcision. Se bilag 27 for eksempler på falske positiver. Omvendt er den relativt høje sensitivitet et tegn på, at AI-redskabet ikke medfører særligt meget falske negativer.

Det kan ses i tallene i tabel 5, at jo tydeligere og mere afgrænset et genstandsfelt er, jo bedre er AI-redskabet til at klassificere dokumenter. Det er lettere for en sprogmodel at klassificere, om noget omhandler AI, som for techvirksomheden, da det er et konceptuelt stramt afgrænset område, kontra at klassificere om et dokument om fjernvarme falder under energipolitik, som for Dansk Standard Energi. Det er interessant, at Danske Gymnasier, som er den eneste, der fast læser monitoreringen, ikke scorer særligt højt. Dette kan forklares ved størrelsen af deres genstandsfelt, som uddybes herunder.

8.1.2.5 Betydningen af genstandsfeltets størrelse

Genstandsfeltets størrelse betyder, at der er forskel på mængden af dokumenter for de forskellige områder. For eksempel er Danske Gymnasiers og Dansk Standard Energis præcision ved negativ tolkning henholdsvis 38% og 40%, hvilket er næsten identisk. Hvis man i stedet ser på forskellen i absolutte tal, er der 2 falske positiver ved Danske Gymnasier og 17 falske positiver ved Dansk Standard Energi. Mængden af støj, som falske positiver repræsenterer, er dermed større hos Dansk Standard Energi end hos Danske Gymnasier, hvilket gør det mere forstyrrende for Dansk Standard Energi i forhold til Danske Gymnasier. Denne støj går i bund og grund ud over brugeroplevelsen og dermed nytten af AI-redskabet.

8.1.2.6 Betydningen af kapaciteten til at optage nyheder fra Folketinget

I den manuelle tildeling af labels er der åbnet døre for at man kan tolke resultaterne enten positivt eller negativt. Siden Danske Gymnasier har interesse i at være ajour med hvad der

sker i Folketinget kombineret med at de har et stærkt sekretariatet og et lille genstandsfelt, er de mere tilbøjelige til at ligge i den positive tolkning, da ikke er særligt forstyrret af mængden af falske positiver (DG1; DG2). Omvendt er SJ&K en lille organisation med tre forskellige klienter og en mindre interesse i at vide, hvad der sker i Folketinget, og derfor er de tilbøjelige til at tolke resultaterne negativt (SJ&K1).

8.1.2.8 Konklusion på kasse 2

Denne kasse i programteorien omhandler, at der identificeres potentielt relevante dokumenter af AI-redskabet. Universelt set er denne kasse ikke verificeret.

SJ&K har ikke prioriteret at følge politiske processer og tilhørende dokumenter i tæt grad givet deres lille organisation. Kun techvirksomhedsprompten, som er én ud af tre prompts, giver gode resultater.

Zero Carbon Shipping har ikke stor interesse i dansk politik, hvilket stiller ekstra høje krav til, hvad der er relevant. Selvom AI-redskabets klassifikationer scorer næstbedst ved Zero Carbon Shipping, er det ikke nok til at opfylde deres behov.

Dansk Standards prompts er brede og genstandsfeltet er stort, hvilket går ud over redskabets ydeevne til at klassificere dokumenter, hvorfor AI-redskabet ikke er god til at klassificere dokumenter til dem, da der er for mange falske positiver.

Danske Gymnasier har derimod stor interesse i at være ajour med hvad der sker i Folketinget og ungdomsuddannelser som genstandsfelt er lille. Derfor, selvom resultaterne fra AI-redskabet ikke er tæt på at være 100% korrekte for Danske Gymnasier, er det stadig en nyttig klassifikation for Danske Gymnasier, da de ikke bliver for forstyrret af falske positiver (DG1; DG2).

Derfor er denne kasse **ikke verificeret**, men der er en delvis verificering af promptsene for Zero Carbon Shipping og Techvirksomheden i den positive tolkning, da de scorer højest på kombinationen af præcision og sensitivitet. Kassen er også **tæt på at være fuldt verificeret** for Danske Gymnasier, når man tager deres behov og genstandsfelt i betragtning. Derfor kan man gå videre til næste skridt i programteorien.

8.1.3 Kasse 3: Relevante dokumenter bliver leveret i organisationens indbakke

Ligesom kasse 1 er denne kasse ikke umiddelbart interessant, men den er helt afgørende. Hvis organisationerne ikke har mulighed for at modtage og læse monitoreringsrapporten, underminerer det hele formålet med AI-redskabet. AI-redskabet fungerer automatisk, uden nogen form for afvigelser. Det forekommer dog, at distributionen af monitoreringsrapporten kan blive forsinket til fredag eftermiddag - hvor der er en tendens til ikke at blive læst mange mails - da processen med at køre AI-redskabet, som skal håndtere otte forskellige prompts hver fredag, er tidskrævende. Ikke desto mindre **er denne kasse verificeret.**

8.1.4 Kasse 4: Organisationer orienterer sig i dokumenter fra monitorerings-rapporten

Organisationerne har nu modtaget deres monitoreringsrapporter i indbakken. Men bliver rapporterne faktisk læst? Det afhænger af organisationen. Danske Gymnasier læser monitoreringsrapporten fast, mens det ikke er tilfældet for de øvrige tre.

I SJ&K bliver monitoreringsrapporten ikke læst fast, fordi SJ&K som organisation er for lille, og det er ressourcekrævende at læse monitoreringsrapporten, selvom indholdet er interessant for dem (SJ&K1). Zero Carbon Shipping udtrykker den samme observation: "Der er bare for meget, og det er uoverskueligt" (ZCS2). Dansk Standard indikerer også, at monitoreringsrapporternes omfang er en barriere. De bemærker, at "det bliver for uoverskueligt… Du får et dokument på 50 sider" (DS1). De tre organisationers svar ligger i tråd med fundene fra kasse 2 om, at der er for mange falske positiver, hvilket gør monitoreringsrapporten for lang og uoverskuelig at læse.

Hos Danske Gymnasier opleves informationsmængden anderledes. De finder ikke informationen direkte irrelevant, men tilkendegiver stadig, at der er dokumenter i monitoreringsrapporten, som ligger for langt ud i periferien af, hvad der er interessant for dem (DG2; DG3).

Omfanget af monitoreringsrapporterne er også passende for Danske Gymnasier af to grunde: For det første modtager de monitoreringsrapporterne dagligt i stedet for ugentligt, hvorfor de er mindre omfattende, da der er blevet akkumuleret færre dokumenter (DG2). For det andet er gymnasieområdet et mindre område end for eksempel energi eller byggeri, som Dansk

Standard er interesseret i. Selvom ikke alt indhold i monitoreringsrapporten er relevant for Danske Gymnasier, er mængden af information håndterbar (DG2). Danske Gymnasier bemærker at "der er ikke så meget støj" og tilføjer, at enkelte irrelevante informationer ikke udgør et stort problem (DG1).

Således kan det bekræftes, at Danske Gymnasier aktivt orienterer sig i monitoreringsrapporten, men det er ikke tilfældet for de andre organisationer. En universel konklusion for denne kasse i programteorien er derfor, at denne kasse **ikke kan verificeres.** Dog orienterer Danske Gymnasier sig fast i monitoreringsrapporten, hvorfor **denne kasse er verificeret ud fra Danske Gymnasiers kontekst**. Jeg finder det derfor relevant at udfylde resten af programteorien for at få mere viden til at besvare problemformuleringen. Det sker dog kun med svar fra Danske Gymnasier.

Efter kasse 4 deles programteorien op i to spor. Det er helt lavpraktisk; hvis der er indhold i monitoreringsrapporten fordi der er blevet klassificeret relevante dokumenter eller hvis monitoreringsrapporten er tom fordi der ikke er blevet klassificeret nogle relevante dokumenter eller kun består af falske positiver.

8.1.5 Der tages en aktiv beslutning om hvordan informationen fra relevante dokumenter skal bruges

I denne kasse bliver der taget direkte stilling til et dokument.

Danske Gymnasier gør det klart, at det er sjældent, at de reagerer direkte på baggrund af den information, de modtager, hvilket også inkluderer den information de får fra deres eksisterende arbejdsgang med at få viden via nyhedsbrevet fra Børne- og Undervisningsudvalget (DG3). Oftest går de ikke længere end til at dele informationen i organisationen. Når der er blevet akkumuleret nok viden, diskuterer de i sekretariatet, om der burde tages specifikke skridt eller handlinger baseret på de indkomne oplysninger, men de har ikke taget stilling til en politisk sag direkte baseret på AI-redskabets monitoreringsrapport (DG3).

Der er dog et lavpraktisk eksempel på hvordan et dokument fra monitoreringen er blevet benyttet. Der blev sendt et beslutningsforslag, som blev stillet i regi af digitaliseringsudvalget om en opdatering af tidsplan, hvor Danske Gymnasier siger: “Og den opdatering af tidsplan,

og den dagsordensættelse af, hvornår der skal behandles i folketingssalen, er jeg ikke sikker på, at jeg havde set, hvis det ikke var for din politiske monitorering." og at der derfor er blevet opsat en kalender-reminder (DG3). Det er dermed et helt konkret eksempel på, hvordan Danske Gymnasier aktivt har brugt AI-redskabet, selvom det er en lille ting at sætte en reminder i kalenderen. Derfra er det uklart, om der bliver ageret på den viden om dagsordenen, da datoerne fra dagsordenen ligger efter interviewserien er afsluttet.

Dermed kan denne kasse **ikke direkte verificeres**, da det er sjældent at Danske Gymnasier reagerer direkte på informationer de modtager fra nyhedsbreve jf deres nuværende arbejdsgang og AI-redskabets monitoreringsrapporter, da det er et samsurium at ting der gør, at de tager stilling og agerer på ting (DG2).

Der er dog positive tendenser, der indikerer en potentiel verificering af denne kasse: AI-redskabet har ikke været implementeret længe nok til, at der er kommet svar på de spørgsmål, der er blevet stillet. Danske Gymnasier orienterer sig i monitoreringsrapporten, hvilket betyder, at AI-redskabet er en del af det samsurium, der bidrager til at tage en aktiv beslutning. Danske Gymnasier siger: "Jeg tror, at på et eller andet tidspunkt, så vil der komme et eller andet, som vi ikke ville have set, fordi vi ikke abonnerer på det udvalg, og så vil vi gøre et eller andet med det." (DG2). Dermed er det sandsynligt, at denne kasse **potentielt kunne verificeres**, hvis AI-redskabet havde været implementeret i længere tid.

8.1.6 Kasse 6: Beslutningen iværksættes i et givet projekt eller på en anden måde

Her bliver en sag, der bliver læst fra AI-redskabets output, faktisk handlet på via konkrete arbejdsprocesser. Selvom der blev sat en kalender-reminder op baseret på et dokument fra monitoreringsrapporten, er der ikke sket en direkte public affairs handling på det endnu (DG3). Denne kasse kan derfor **ikke verificeres** per definition, da den forrige kasse ikke er fuldt ud verificeret (Dahler-Larsen, 2018)**.**

8.1.7 Kasse 7: Viden inkorporeres i public affairs arbejdet som baggrundsviden

I denne kasse bliver et relevant dokument læst og forstået, men der bliver ikke direkte ageret på den.

Danske Gymnasier ser værdi i at være opdateret på politiske bevægelser i Folketinget via AI-redskabet (DG1; DG2; DG3). Dette understøttes af, at de ikke nødvendigvis bliver irriteret over indholdet i monitoreringsrapporten i konteksten af for meget støj eller irrelevant information. De sender ofte rapporten videre til relevante kollegaer afhængig af sagens karakter, hvilket kan dreje sig om alt fra datasikkerhed til juridiske spørgsmål om personaleret. Dette viser en proces, hvor information selektivt distribueres internt (DG1). Eksempler på dette er, at monitoreringsrapporten den 18. marts indeholdt to relevante dokumenter: ét om Chromebook-sagen i Digitaliseringsudvalget og ét om arbejdsmiljø og BPA i Beskæftigelsesudvalget. De fremhæver, hvordan sådanne informationer uden AI-redskabet kunne være gået ubemærket hen (DG1). For eksempel blev information om Chromebook-sagen sendt videre til Uddannelses- og databeskyttelseskonsulenten, der fandt det relevant, mens oplysninger om arbejdsmiljø blev sendt til Analyse- og digitaliseringskonsulent, der kunne finde det relevant i sit arbejde (DG1).

Danske Gymnasier påpeger også, at de primært har brugt informationerne til at øge deres vidensniveau (DG1). Dette er særligt relevant, hvis der kommer svar på specifikke spørgsmål fra kilder, som Danske Gymnasier ikke normalt selv følger så tæt, som kan afsløre divergerende holdninger mellem ministerier. Et eksempel er et samråd fra Udvalget for Landistrikter og Øer, hvor ministeren fik et spørgsmål omhandlende udkantsgymnasier. Danske Gymnasier siger: "når vi så får svar på det spørgsmål, så vil vi jo bruge det, hvis der er noget indholdsmæssigt, der er relevant, så vil vi jo bruge det til at orientere vores bestyrelse, også bruge det i vores politiske interessevaretagelse." (DG1). Der nåede ikke at komme svar på dette spørgsmål inden interviewserien med Danske Gymnasier blev afsluttet.

Danske Gymnasier påpeger, at anvendelsen af disse informationer strækker sig ud over interne medarbejdere; de integreres også i kommunikationen til bestyrelsen. Her bliver de brugt til at informere om vigtige emner som elevfordeling og nyeste ændringer i taxameter. Disse opdateringer bruges til at tage stilling til, om der skal foretages ændringer i deres tilgang eller politik, hvilket viser, at information fra AI-redskabet er mere end blot interne opdateringer (DG2).

Disse eksempler demonstrerer, hvordan Danske Gymnasier håndterer og anvender information til både intern koordination og orientering til deres bestyrelse.

I det sidste interview med Danske Gymnasier, siger de, at AI-redskabet ikke har bragt en politisk sag til deres kendskab, som de ikke var opmærksomme på før. Danske Gymnasier siger, at der er nogle detaljer omkring nogle forløb, som de er blevet opmærksomme på (DG3). Det er altså i en lille skala, som AI-redskabet bidrager som et nyttigt redskab til at indsamle baggrundsviden.

Denne kasse **kan derfor verificeres** for Danske Gymnasier.

8.1.8 Kasse 8: Organisationer er ikke bekymret for at overse relevant information

Denne kasse omhandler, at når AI-redskabet har lavet klassifikationen og sendt monitoreringsrapporten, så er organisationerne ikke bekymrede for, at de overser relevant information.

Det er dog en udfordring at verificere denne kasse, da det er svært at måle fraværet af noget, der enten ikke eksisterer eller ikke er fanget i organisationernes opmærksomhed. Derfor afhænger det af organisationernes tillid til, at AI-redskabet effektivt identificerer alle relevante dokumenter. Tillid bruges dermed som en proxy til at vurdere, om organisationerne ikke er bekymrede for at overse relevant information.

De udtrykte til at starte med en vis tillid til AI-redskabet (DG1). Sidenhen, i et efterfølgende interview, er Danske Gymnasier blevet opmærksom på at der har været falske negativer, altså relevante dokumenter, der ikke kom med i monitoreringsrapporten. Dette har medført en vis mistillid hos Danske Gymnasier. Til gengæld har de tillid til AI-redskabet, hvad angår det, der ligger uden for Børne- og Undervisningsudvalgets område (DG3).

Derfor konkluderes det, at denne boks endnu **ikke kan verificeres**, grundet den ikke-universelle tillid til AI-redskabets evne til at fange alle relevante dokumenter.

8.1.9 Kasse 9: Større viden om hvad der sker i Folketinget

Danske Gymnasier værdsætter AI-redskabet og har tillid til, at det vil frembringe relevante dokumenter om, hvad der sker i Folketinget (DG1; DG2; DG3). Dog bemærker de, at AI-redskabet indtil videre ikke har leveret information, som de ikke allerede kunne have opdaget gennem deres eksisterende arbejdsgange, men AI-redskabet har til gengæld leveret

flere detaljer om disse politiske sager (DG3). De understreger dog, at AI-redskabet har potentialet til at opdage informationer, som de måske ville have overset, da det dækker flere udvalg end blot Børne- og Undervisningsudvalget, samt at der allerede er identificeret dokumenter i monitoreringsrapporten, som de har optaget som baggrundsviden (DG1; DG3).

Danske Gymnasier fremhæver, at AI-redskabet systematisk samler information fra forskellige politiske områder, som de normalt ikke ville fokusere på, hvilket giver dem en mere grundig oversigt. Dette ses som en forbedring i forhold til tidligere, hvor de kun modtog nyheder fra Børne- og Undervisningsudvalget (DG2). Det bekræfter, at redskabet gør det både nemmere og mere grundigt at indsamle information fra Folketinget (DG3).

Derfor kan denne kasse **verificeres,** men kun i begrænset omfang. AI-redskabet har potentiale til at give endnu større viden, hvis den havde været implementeret i længere tid. Igen skal det huskes på, at det kun er Danske Gymnasier, kassen er verificeret for.

8.1.10 Kasse 10: Bedre politiske og strategiske valg baseret på mere viden

Danske Gymnasier påpeger, at AI-redskabet ikke har været implementeret længe nok til at de kan drage endelige konklusioner (DG2; DG3).

Direkte adspurgt om hvorvidt de mener, at de er blevet bedre til at træffe politiske og strategiske beslutninger på baggrund af AI-redskabet, svarer Danske Gymnasier, at det vil være for meget at sige (DG2). Det tilføjes, at det er for tidligt at sige noget definitivt, men de vurderer, at over et år kunne der muligvis være 5-10 dokumenter, som de ikke ville have opdaget uden AI-redskabet (DG2). De antyder, at på længere sigt kunne AI-redskabet forbedre deres evne til at øve politisk indflydelse, hvilket måske kunne have været opnået ved at ansætte en ekstra medarbejder til at foretage mere effektiv medieovervågning (DG2).

Derfor kan denne boks **ikke verificeres,** men Danske Gymnasier ser potentiale med tiden.

8.2 Opsamling på evalueringen

Man kan dermed konkludere, når man inkluderer alle fire organisationer, at AI-redskabet **ikke virker jf programteorien**. Der er for mange falske positiver, indholdet skal prioriteres, og mængden af information er for overvældende.

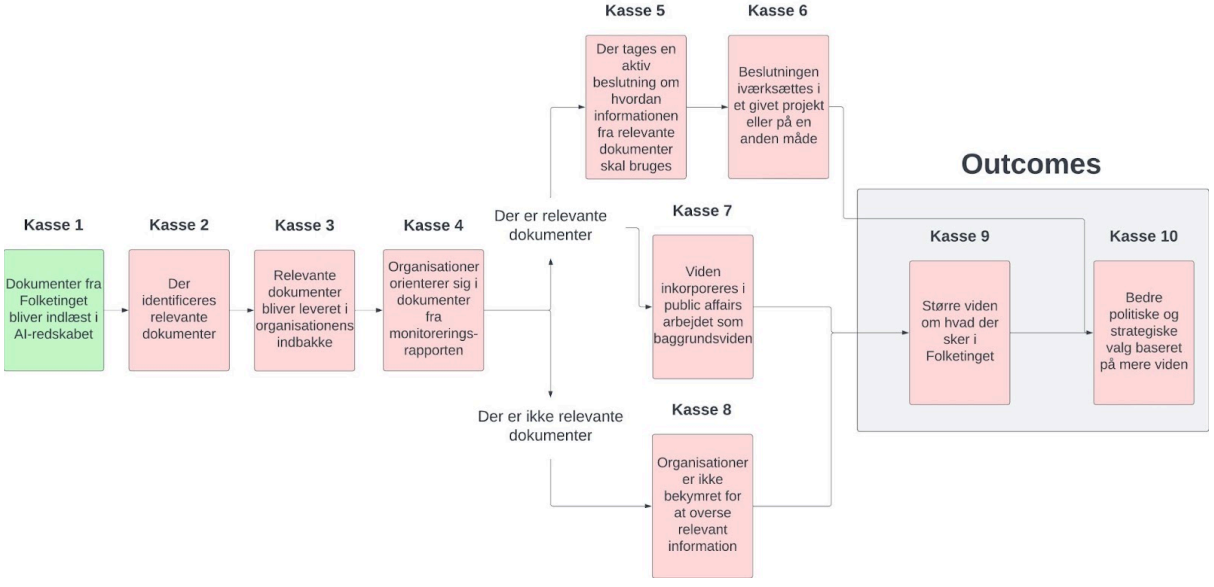
Dog kan man konkludere, at AI-redskabet fungerer til en vis grad hos Danske Gymnasier til at opnå større viden om, hvad der sker i Folketinget. Dette stemmer overens med Pawson og Tilleys pointe om, at konteksten har indflydelse på, hvorvidt en mekanisme er succesfuld (Pawson & Tilley, 1997). Det kan bruges til at forstå hvorfor AI-redskabet til en vis grad fungerer for Danske Gymnasier:

- Genstandsfeltet er snævert, hvorfor promptet er mere snævert defineret.
- De modtager daglige opdateringer via e-mail, hvorfor mængden af information er spredt ud i mere overskuelige mængder.
- De har en stor interesse i at modtage viden fra Folketinget, ligesom Dansk Standard, men mere end SJ&K og Zero Carbon Shipping.
- De er tolerante over for falske positiver (DG2)

Det skal igen gøres klart, at jeg har den største viden om Danske Gymnasier, som har medført, at værktøjet bevidst og ubevidst er designet efter en arbejdsgang, som matcher Danske Gymnasiers arbejdsgang jf. afsnit 7.5. Det spiller naturligvis en rolle, men baseret på de beregnede tal i afsnit 8.1.2, hvad interviews har kortlagt og måden GenAI fungerer, er konklusionen at de 4 specifikke faktorer nævnt ovenfor stadig er gældende.

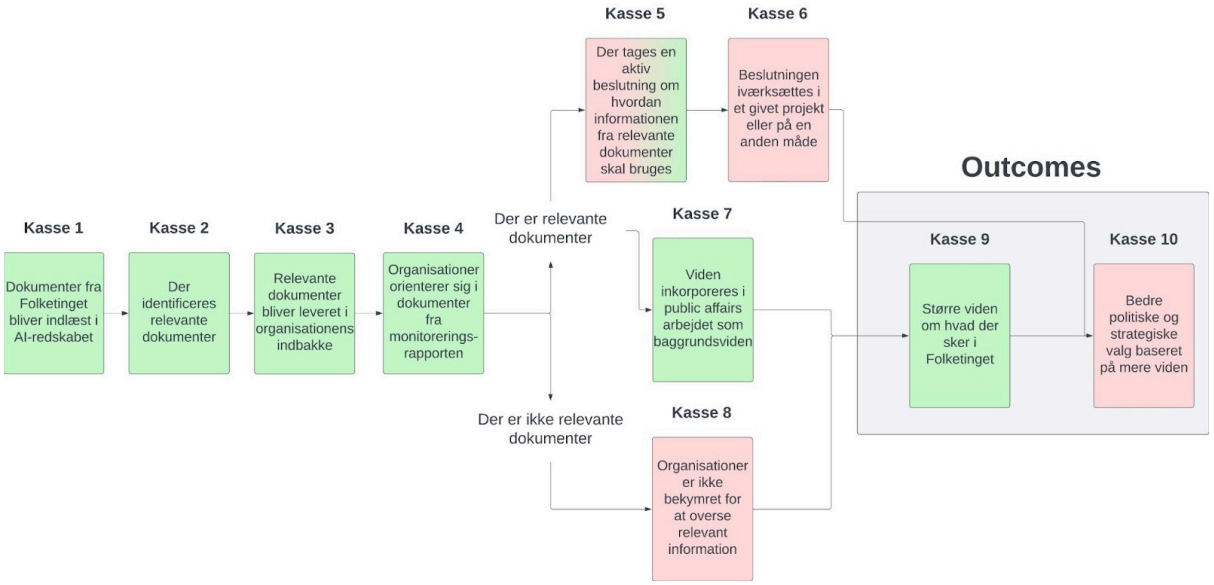
Den fulde konklusion på programteorien kan derfor stilles grafisk op således:

Figur 5: Resultat af evaluering af programteorien for alle 4 organisationer



Konklusionen på programteorien *for Danske Gymnasier* kan stilles grafisk op som nedenfor.

Figur 6: Resultat af evaluering af programteorien for Danske Gymnasier



8.2.1 Teorifejl og implementeringsfejl

Spørgsmålet er, om det er teorifejl, implementeringsfejl eller begge ting (Dahler-Larsen, 2018). Det er generelt muligt at benytte sprogmodeller til at lave tekstklassificeringer (Johansen, 2024), hvorfor det *ikke* er en teorifejl. Ligeledes er det *ikke* en teorifejl om der faktisk er et behov for at klassificere dokumenter fra Folketinget, som alle fire organisationer mener er brugbart og interessant (DG1, DS1, SJ&K1, ZCD1). En potentiel teorifejl er om hvorvidt GenAI er god nok i en dansk kontekst til at lave tekstklassifikationer, men det er udfordrende at konkludere givet at AI-redskabet indeholder implementeringsfejl.

Implementeringensfejlene i AI-redskabet er hvordan AI-redskabet er designet, siden promptsystemet ikke er skarpt nok og medfører for mange falske positiver. 'Off-the-shelves' sprogmodeller med ét prompt per område ikke er gode nok endnu. En anden implementeringsfejl er at monitoreringsrapporten er for uoverskuelig og indeholder for mange falske positiver.

Når resultatet er udeblevet (der er ikke en universel verificering af programteorien) og at AI-redskabet ikke er implementeret korrekt (der er for mange falske positiver og en uoverskuelig monitoreringsrapporten), er anbefalingen fra litteraturen om virkningsevaluering, at man skal justere AI-redskabet og implementeringen af AI-redskabet, men være beredt på at der stadig er usikkerhed forbundet med projektet, siden det ikke er

sikkert, at der i fremtiden vil være anderledes outcomes end det der er belyst i den oprindelige programteori (Dahler-Larsen, 2003). Det diskuteres hvordan man kan justere AI-redskabet fremadrettet i afsnit 9.1.

8.3 Fund ud over programteorien

Som evaluator skal man tage alt med, der er relevant for en evaluering, selvom en indsigt ikke fremgår af i programteorien (Dahler-Larsen, 2018). Disse indsigter fremgår nedenfor:

8.3.1 AI-redskabet skaber stadig værdi og har potentiale

Selvom AI-redskabet ikke fungerer i overensstemmelse med programteorien, betyder det ikke, at det ikke skaber værdi. AI-redskabet har demonstreret, at selv et simpelt redskab som dette kan være af værdi, eksempelvis for Danske Gymnasier. De 3 andre organisationer har også udtalt, at løsningen delvist fungerer og har skabt værdi i visse perioder og kontekster (DS4; SJ&K6; ZCS2). Dette styrkes yderligere ved, at Dansk Standard siger, at på den korte tid, som AI-redskabet har været implementeret, har det nået et niveau, der matcher det, som andre udbydere af lignende løsninger har opnået, som har været dyrere og mere komplekse løsninger (DS1).

Zero Carbon Shipping tilføjer, at de ikke tidligere har haft et så omfattende overblik, som AI-redskabet nu tilbyder, så de ser klare fordele ved at bruge det og mener, at det kan tilføre værdi for andre i lignende situationer, hvis brugerfladen bliver bedre og promptet bliver skarpere. Og selvom de ikke læser monitoreringsrapporten fast, har de tilkendegivet, at de direkte er blevet opmærksom på relevante dokumenter, som de ellers ikke havde været opmærksom på og læst disse dokumenter (ZCS2).

SJ&K påpeger, at redskabet giver et godt overblik, men også at det kan være relativt ressourcekrævende at læse monitoreringsrapporten. Derfor er det vigtigt at finde en måde at gøre brugen af AI-redskabet mindre ressourcekrævende (SJ&K1).

Selvom programteorien ikke er verificeret fuldt ud, viser det, at grundstrukturen er på plads og videreudvikling af AI-redskabet potentielt kan medføre bedre resultater, siden AI-redskabets brug af GPT4 er meget basal, hvorfor der potentielt er mange forbedringer tilgængelige. Det diskuteres i afsnit 9.1.

8.3.2 Behov for prioritering af indholdet i monitoreringsrapporten

SJ&K, Zero Carbon Shipping og Dansk Standarder læser ikke monitoreringsrapporten fast, fordi den er for lang og uoverskuelig. De foreslår alle sammen at der sker en prioritering af indholdet i monitoreringsrapporten, så dokumenterne bliver rangeret efter niveau af relevans (SJ&K1, ZCS2, DS2). Danske Gymnasier læser monitoreringsrapporten fast, men ønsker også at der sker en prioritering eller sortering af dokumenterne i rapporten (DS3).

Danske Gymnasier pointerer også, at der foregår en intern prioritering og sortering af, hvad der skal sendes videre i organisationen fra monitoreringsrapporten, men at sorteringen er af anden karakter end hvad de plejer at gøre ifm. nyhedsbrevene. Det skal forstås som, at nyhedsbrevet fra Børne- og Undervisningsministeriet også indeholder dokumenter om folkeskolen, som ikke er relevante for Danske Gymnasier. Med AI-redskabet er det i stedet en sortering af falske positiver, der ikke har direkte relevans for Danske Gymnasiers arbejde. (DG1; DG3).

8.4 Bivirkninger jf programteorien

Programteorien giver mulighed for at undersøge eventuelle bivirkninger ved at anvende AI-redskabet (Dahler-Larsen, 2018). To bivirkninger, der er blevet identificeret, er en mere struktureret tilgang til håndtering af data og et potentiale for en database.

8.4.1 Struktureret tilgang til data

Danske Gymnasier fremhæver, at den systematiske behandling af data er positiv, forstået på den måde, at AI-redskabet systematisk gennemgår alle dokumenter fra Folketinget og det derfor ikke er overladt til tilfældighederne at opdage dokumenter, der ligger uden for deres genstandsfelt. Dette mindsker risikoen for at overse relevante informationer (DG3).

SJ&K påpeger, at AI-redskabet har ført til en øget bevidsthed inden for deres organisation omkring beslutningsprocesserne i Folketinget, da de nu modtager data fra Folketinget struktureret. Det har været en katalysator for at sikre, at organisationen holder sig opdateret med aktiviteterne i Folketinget, hvilket har haft både organisatorisk og psykologisk påvirkning. Dette har kvalificeret og professionaliseret den måde, organisationen diskuterer politiske emner på, og har generelt forbedret de politiske samtaler internt. SJ&K mener, at

engagementet i at bruge AI-redskabet, på trods af at de ikke læser det fast, har motiveret medarbejderne til at gøre en ekstra indsats og diskutere politiske beslutningsprocesser på en mere kvalificeret måde. Derfor omtaler SJ&K AI-redskabet som et "sociologisk professionaliseringsværktøj" (SJ&K1).

Denne bivirkning vedrørende en mere struktureret tilgang til data er interessant da GenAIs kapacitet til at strukturere data kan bruges i public affairs til at tilgå de enorme mængder kvalitativ data, der er i public affairs feltet og give en øget systematik til tilgangen af data.

8.4.2 AI-redskabet som database

En anden bivirkning, der blev identificeret, var, at SJ&K tilkendegav, at de ikke læste monitoreringsrapporterne fast, men de fandt stadig nytte ved AI-redskabet, hvilket virkede usammenhængende (SJ&K1). Da der blev spurgt ind til det, viste det sig, at en utilsigtet bivirkning af AI-redskabet opstod i forbindelse med SJ&Ks arbejde for en socialorganisation. Her udnyttede SJ&K de informationer, som AI-redskabet havde frembragt i tidligere monitoreringsrapporter, til at danne et overblik over politiske dokumenter inden for socialområdet. Ved retrospektivt at gennemgå gamle dokumenter identificerede SJ&K politiske bevægelser og tendenser, som var relevante for socialorganisationens interesser (SJ&K1,5).

Dette fremhæver en væsentlig anvendelse af AI-redskabet, som ikke var forudset: evnen til at danne et overblik over tidligere politiske dokumenter, der kan bidrage til strategisk rådgivning i public affairs. Denne bivirkning understreger AI-redskabets potentiale ikke kun som et værktøj til løbende overvågning, men også som en ressource til at mappe politiske dokumenter og det politiske landskab, hvilket er værdifuldt (Esbensen, 2007; Hegelund & Mose, 2013). Hvordan dette potentiale kan indfries, diskuteres i afsnit 9.2.