Tech-giganter og demokrati: Kunstig intelligens, misinformation og digitale platforme

Mads Fuglsang Hove, Syddansk Universitet
Rebecca Adler-Nissen, Københavns Universitet
Anja Bechmann, Aarhus Universitet
Claes H. de Vreese, Syddansk Universitet
Frederik Hjorth, Københavns Universitet
Yevgeniy Golovchenko, Københavns Universitet

Acknowledgements: Tak til Sofia Tang fra Københavns Universitet for hjælp til at generere billeder til eksperimentet.

Executive summary

Denne rapport omhandler tech-giganternes indflydelse på det danske demokrati, den offentlige samtale og den sociale sammenhængskraft. Den undersøger særligt fænomenet misinformation samt de effekter som sociale medier har på politisk polarisering og trivsel i Danmark. Overordnet påviser rapporten, at tech-giganterne har stor indflydelse på den demokratiske samtale i Danmark og potentielt endnu større i takt med udviklingen af generativ kunstig intelligens. Der er imidlertid fortsat mangel på systematiske undersøgelser relateret til tech-giganter, demokrati og sammenhængskraft – særligt i Danmark. Rapporten er udarbejdet som en del af Medieaftalen for 2023-2026 og er den første i en række årlige rapporter.

Nøglefund

- Danskerne klarer sig ikke væsentlig bedre end tilfældige gæt, når de skal identificere, hvorvidt billeder er skabt med generativ AI eller er ægte billeder. Der er dog forskelle i evnerne blandt danskerne, hvor bl.a. yngre, kvinder samt dem som er mere analytisk tænkende og de, som har prøvet kræfter med generativ AI før, er bedre til at vurdere ægtheden af et billede. Vores undersøgelse har betydning for, hvordan vi skal forstå danskernes evner til at identificere eventuel misinformation skabt med generativ kunstig intelligens og hvilke samfundsgrupper, der er mest udsatte i den henseende.
- Danskerne bliver mere bange for, om de kan stole på, hvad de ser online, og om misinformation påvirker politiske valg, når de får at vide, at misinformation er et stort samfundsmæssigt problem. Men vores undersøgelse peger også på, at danskerne godt kan håndtere nuanceret information uden demokratisk negative følgevirkninger.
- Danskerne er generelt, tilhængere af, at de sociale medier modererer og sletter indhold, der er stødende, ulovligt eller farligt. Danskerne støtter i overvejende grad de eksisterende retningslinjer for indholdsmoderation. Der er dog særligt indikationer på kulturelt betingede forskellige opfattelser af, hvad danskere mener bør slettes eller bevares, og de (typisk amerikanske) sociale mediers retningslinjer herfor, f.eks. ift. topløshed.
- Der er meget, der tyder på, at mængden af misinformation på sociale medier er af mere begrænset omfang end man ofte hører, mens der er større usikkerhed om, hvilken betydning misinformationen har mere generelt på samfundet. Det skyldes ikke mindst, at misinformation kan komme fra mange forskellige mere eller mindre vægtige aktører, være skabt til at vække engagement og være svær at genkende, enten fordi der er brugt sofistikerede teknikker, eller fordi det er integreret i ellers faktuelt korrekt information.
- Vi mangler fortsat mere viden om – særligt danskernes – adfærd på sociale medier. Det gælder både, i hvilket omfang de eksponeres og interagerer med misinformation og politisk polariserende indhold, samt mængden og typen af indhold, der er problematisk for danskernes og særligt børn og unges trivsel.

Vigtige implikationer

- Manglende adgang til data fra tech-giganter, heriblandt data om brugernes adfærd, er den største hæmsko i forhold til at undersøge tech-giganternes betydning for samfundet. Langt de fleste forskningsmæssige uenigheder og uafklarede spørgsmål bunder i, at der sjældent er adgang til de nødvendige data for at kunne svare på de komplekse spørgsmål, samfundet er mest interesserede i. Dataadgang er derfor

afgørende for at etablere et præcist billede af udfordringerne med misinformation, der hverken over- eller underdriver problemernes samfundsmæssige omfang.

- Danske myndigheder og lovgivere bør arbejde for at sikre, at bedre muligheder for dataadgang og lignende fra bl.a. EU-forordningen om digitale tjenester (Digital Services Act) kan anvendes af forskere, myndigheder og civilsamfund. Det kræver vedholdenhed og ressourcer, der både kan hjælpe med at sikre adgang til tilstrækkelige data samt støtte aktørerne juridisk fra sag til sag, så økonomi og ulige magtforhold ikke afholder aktører fra at søge og arbejde med sådanne data.
- Alle kan blive manipuleret af misinformation skabt med generativ AI, men det er bestemte samfundsgrupper, der risikerer at være i størst fare for at falde for misinformation. Bedre uddannelse og teknologiforståelse blandt de mest udsatte grupper kan muligvis bidrage til at mindske mængden af danskere, der har sværest ved at navigere i en verden med meget generativ AI skabt information.
- Det er vigtigt, at toneangivende aktører som politikere, myndigheder og medier forholder sig nuanceret til fænomener som misinformation uden at underkende eller overdrive problemets omfang. Der er meget, der tyder på, at danskerne kan navigere i et nuanceret og transparent informationsmiljø, og at denne måde at kommunikere på samtidig undgår de negative effekter på eksempelvis opbakning til demokrati som styreform.

Indholdsfortegnelse

Forord 6
Tech-giganternes indflydelse på samfundet 7
Tech-giganternes forretningsmodeller 7
Rammerne for den offentlige samtale på sociale medier 9
Aktive brugere på sociale medier og sociale mediers annonceindtægter 10
Vurderinger 14
Hvad ved vi om sociale mediers betydning for misinformation, politisk polarisering og trivsel? 15
Misinformation på sociale medier 16
Sociale mediers betydning for spredning af misinformation 17
Udbredelse og tro på misinformation 18
Effekten af interventioner mod misinformation 20
Politisk polarisering 23
Sociale mediers betydning for trivsel 24
Fortsat ubesvarede spørgsmål og forslag til forskningsdesigns 26
Fortsat ubesvarede spørgsmål 26
Misinformation 26
Politisk polarisering 28
Trivsel 29
Forslag til fremtidige temaer og forskningsdesigns 29
Forskningsdesign I: Danskernes adfærd med misinformation online 29
Forskningsdesign II: Politisk polarisering på sociale medier i Danmark 30
Forskningsdesign III: Social mediers betydning for danske børn og unges trivsel 31
Begrænsninger og ressourcekrav for de tre forskningsdesigns 32
Ny undersøgelse af danskernes syn på og evner ift. misinformation og generativ AI 33
Holdninger til indholdsmoderation og regulering på sociale medier 33
Danskernes evner til at identificere AI-genereret indhold 36
Betydningen af opmærksomheden rettet mod misinformation 41
Vurderinger 44
Undersøgelsens opdrag og organisering 46
Litteraturliste 47
Bilag 54
Bilag 1A: Indsamling af aktivitets- og annoncedata fra Facebook 54

Bilag 2A: Identifikation og systematisering af forskningslitteratur 55
Bilag 3A: Holdninger til indholdsmoderation på Facebook 56
Bilag 3B: Holdninger til reguleringsanbefalinger 57
Bilag 3C: Design af AI-billede genkendelse eksperiment 58
Bilag 3D: Specificering af regressionsmodeller af AI-billede eksperiment 63
Bilag 3E: Design af misinformationsopmærksomhed eksperiment 63

Forord

Graden af optimisme forbundet med internettet har bevæget sig over bakke og dale. Fra dets spæde start oplevede nyskabende services og fællesskaber som Ebay og MySpace stor interesse. Senere nåede optimismen dens højderyg med det arabiske forår, Occupy Wall Street-bevægelsen og valget af Barack Obama som amerikansk præsident. Demokratisering, inklusion af stemmer oftest udeladt fra debatten og en mere direkte relation mellem politiker og borger var nogle af de elementer, der skabte en opfattelse af, at tech-giganter, herunder sociale medier, stod i kernen af samfundet og demokratiet. Vi skulle alle på Facebook, sige vores mening og udvide vores netværk.

Men løber man stærkt, risikerer ting at gå i stykker. Og stærkt har alle platformene løbet. Facebook, Google, Instagram, Twitter, YouTube, TikTok og flere. Med skandaler som Cambridge Analytica, russiske forsøg på at påvirke demokratiske valg, konspirationsnetværk og senest frygten for effekterne på børn, unge og voksnes trivsel befinder vi os nu i pessimismens dal, når det gælder synet på tech-giganter. Skandaler, der naturligt har medført et samfundsmæssigt fokus på at skabe et solidt, systematisk og videnskabeligt grundlag for, hvordan vi skal forstå tech-giganternes indflydelse på samfundet. For på den måde at vide, om der er grund til at være bekymret, og hvad vi i så fald bør gøre ved det.

Rapporten her er den første i en række, der som led i Medieaftalen for 2023-2026 giver indblik i centrale emner, når det gælder tech-giganters indflydelse på dansk demokrati, sammenhængskraft og trivsel. Overskriften er bred, men indsigterne er specifikke. Denne rapport fokuserer på spørgsmålet om misinformation på sociale medier, hvad forskningen siger om det, og hvordan danskerne forholder sig til det. Vi gennemgår, hvad vi ved, og hvad vi ikke ved fra forskningen, og indsamler selv data fra et repræsentativt udsnit af den danske befolkning med et spørgeskema og to eksperimenter, der viser nyt om, hvordan danskerne forholder sig til information på sociale medier. Derudover har rapporten kortere afsnit, der handler om emnerne politisk polarisering og trivsel, og som peger fremad mod fremtidige rapporter.

Konkret indeholder rapporten fire dele. For det første optegnes nogle af de større strukturelle strømninger, der giver tech-giganterne liv og udfordrer eksisterende samfundsstrukturer. For det andet viser vi, hvad vi ved fra forskningslitteraturen med særligt fokus på misinformation samt sociale mediers betydning for politisk polarisering og trivsel. For det tredje peger vi på centrale spørgsmål, forskningen endnu ikke i tilstrækkelig omfang har afdækket, samt forslag til, hvordan man kan svare på nogle af de spørgsmål. For det fjerde spørger og tester vi danskerne i en række centrale emner, som f.eks. deres holdninger til indholdsmoderation, deres evne til at identificere AI-genereret indhold og hvordan de bliver påvirket af måden, farerne ved misinformation omtales på.

Rapporten er udarbejdet i et samarbejde mellem forskere fra tre interdisciplinære forskningsmiljøer: Digital Democracy Centre (Syddansk Universitet), Copenhagen Center for Social Data Science (Københavns Universitet) og Center for Digital Social Research (Aarhus Universitet).

Tech-giganternes indflydelse på samfundet

Tech-giganternes indflydelse på samfundet spænder over et kæmpe område, hvilket umuligt kan dækkes her. Vi fokuserer derfor på tre elementer, der er særdeles vigtige, når det gælder, hvorfor, hvordan og til hvem misinformation spredes på sociale medier. Konkret vil vi beskrive strømninger og udfordringer relateret til tech-giganternes forretningsmodeller, hvordan sociale medier skaber rammerne for den offentlige samtale samt antallet af brugere og annonceindtjening.

Tech-giganter er en samlebetegnelse for virksomheder, som i varierende grad bygger deres forretningsmodel på at indsamle enorme mængder data om brugerne. Den data bruger og videreformidler tech-giganterne til bl.a. tredjeparter, der f.eks. anvender oplysningerne til målrettede reklamer, til at optimere deres egen forretning og til at fastholde brugerne (Erhvervsministeriet, 2024). Hvor betegnelsen også normalt omfatter platforme som Amazon og Uber, bruger vi i rapporten her primært betegnelsen om platforme, som spiller en udbredt rolle i den demokratiske samtale såsom Facebook og Google.

Andre relevante spørgsmål må derfor beskrives og undersøges andetsteds, såsom demokratisk og politisk deltagelse, algoritmiske biases, der forstærker eksisterende uligheder i samfundet, teknologiens tårnhøje energiforbrug, der modsætter sig kampen mod klimaforandringer, og arbejdsforhold, hvor svage samfundsgrupper risikerer at blive udnyttet.

Tech-giganternes forretningsmodeller

Sociale medier er i den grad blevet en vigtig del af rygraden i den offentlige debat. Historier *breakes* og indlæg i debatten foregår på sociale medier, og sågar hele TV udsendelser bygges op om f.eks., hvilke *tweets* Donald Trump senest har sendt. Mange borgere bruger også sociale medier som nyhedskanal. For eksempel siger omtrent 50 procent af unge amerikanske TikTok brugere, at de bruger platformen til at holde sig opdateret om politik (Pew, 2024). Og selvom de fleste måske mest er interesserede i underholdende elementer på sociale medier, så bliver de fortsat eksponeret for nyheder og politik, der øger deres politiske viden og engagement (Nanz & Matthes, 2022).

Men intet kommer ud af ingenting. For selvom adgangen til mange tech-giganter som Google og Facebook umiddelbart er gratis, så tjener firmaerne bag rigtig mange penge. Indtjeningen drives i de fleste tilfælde af de data, platformene indsamler om deres brugere. Data, som annoncører kan bruge til at målrette annoncer til præcis det segment, de er interesserede i. For adgang til sociale medier, søgemaskiner og lignende betaler vi derfor ikke med en månedlig overførsel, som man kender det fra f.eks. et Netflix-abonnement, men i stedet med de digitale spor, vi efterlader.

Forretningsmodellen medfører også, at mere vil have mere. Jo længere tid, en bruger er på platformen, desto flere annonceindtægter og data kan platformen hente. Data, der igen kan gøres tilgængelig for annoncører, og som kan bruges til at vise mere målrettet indhold til den

enkelte bruger, som så befinder sig længere tid på platformen, fordi indholdet fastholder brugeren.

Det beskrevne loop kaldes også for *the hype machine* og er ofte, hvor pegefingeren ender, når dårligdomme diskuteres (Aral, 2021). Hvis vi kun får serveret det indhold på sociale medier, som vi godt kan lide, påpeger flere risikoen for ekkokamre, filterbobler og polarisering (Sunstein, 2001; Pariser, 2011). Hvor dygtige Facebook og andre tech-giganter er til at bruge den data, de har om os, om det er en god forretning for annoncørerne, og hvorvidt det hele medfører ekkokamre, ligger alt sammen udover denne rapports sigte. Senere afsnit vil i stedet komme nærmere ind på, hvad forskningen siger om sociale mediers betydning for misinformation, polarisering og trivsel.

Forretningsmodellerne og udformningen af platformene anses dog ikke for isoleret at handle om et produkt på linje med andet, vi kan handle online. Tech-giganterne er gået ind i hjertet af samfundet og påvirker offentlige institutioner, økonomiske transaktioner og sociale og kulturelle praksisser (van Dijck et al., 2018). Herfra bliver de langsomt en del af eksisterende samfundsinstitutioner, og tvinger regeringer verden over til en tilpasning af de grundlæggende juridiske og demokratiske rammer for samfundet.

En af de veje, hvorved tech-giganterne sniger sig ind, er ved at promovere deres tjenester som offentlige goder (van Dijck et al., 2018). Strategien er velkendt, men den er imidlertid langt mere potent for tech-giganter. Selvdiagnosticering, alternative behandlinger, private uddannelsestilbud og læringsressourcer er blot et udpluk af et hav af tjenester, samfundet er nødt til at forholde sig til. Et eksempel på problematikken i Danmark er den såkaldte Chromebook-sag, hvor Datatilsynet vurderede, at der ikke er lovhjemmel til at videregive personoplysninger til Google. Landets skoler måtte derfor i en periode leve i uvished om, hvordan undervisningen skulle gennemføres, inden KL i sommeren 2024 indgik en aftale med Google, som man forventer vil leve op til kravene fra Datatilsynet. Tjenester fra Google, Microsoft, Amazon og andre er således blevet afgørende for at den offentlige sektor kan fungere.

Vi ser dog skridt på vejen mod retlige rammer. EU-lovgivning, særligt forordningen om digitale tjenester (DSA) og forordningen om digitale markeder (DMA), sikrer f.eks. mindreårige mod at blive målrettet med personaliserede annoncer, ligesom målretning baseret på sensitive data såsom seksuelle præferencer og religiøs overbevisning er forbudt. Ligeledes regulerer AI-forordningen anvendelsen af kunstig intelligens, der er udbredt og i stor fokus hos tech-giganterne. Således forbydes brugen af kunstig intelligens i tilfælde, hvor det indebærer en uacceptabel stor risiko, som f.eks. at indsamle ansigtsbilleder fra internettet til at oprette ansigtsgenkendelsesdatabaser.

Udfordringen er imidlertid, at regulering, der er nødvendig for at sikre ét gode, nogle gange svækker et andet. For eksempel er gennemsigtigheden i data vigtig for, at myndighederne kan opspore kriminalitet og terrornetværk. Men gennemsigtighed kolliderer bl.a. med hensynet til brugernes privatliv. Sådanne konflikter bringer ligeledes udfordringer til overfladen, hvor der kan være forskel på (primært amerikanske) selskabers tilgang til offentlige samtaler og danske traditioner. En af disse udfordringer er spørgsmålet om indholdsmoderation, som næste afsnit handler om.

Hvor vi i afsnittet her har talt om tech-giganter, skifter vi nu spor til primært at fokusere på den del af tech-giganter, som er sociale medier. Det er platforme som Facebook, Instagram, X (tidligere Twitter), TikTok og Snapchat, hvor brugerne i høj grad har mulighed for at interagere.

Rammerne for den offentlige samtale på sociale medier

Hvor vi får vores nyheder fra, har ændret sig dramatisk siden fremkomsten af sociale medier. Således er sociale medier en af de mest foretrukne og brugte måder at tilgå nyheder, særligt blandt unge (Baptista & Gradim, 2020). I Danmark er Facebook fortsat den sociale medieplatform, der hyppigst anvendes til nyheder (32%), selv blandt unge (34%), mens også Instagram (19%) og TikTok (15%) anvendes ofte (Newman et al., 2024: 76). Og der er masser af nyheder at tage af. Hver dag sendes der mere end 100 milliarder beskeder på alene Facebooks produkter, og hvert minut uploades der over 500 timers video til YouTube (Morrow et al., 2021).

En af udfordringerne, der følger, er dog, at indholdet på sociale medier – og internettet generelt – i lang tid har haft smag af det vilde vesten. Én blandt mange andre situationer, var i forbindelse med indledelsen af folkedrab på rohingyaer i Myanmar i 2017. Facebook har indrømmet, at dets algoritmer i månederne og årene op til grusomhederne mod rohingyaerne i Myanmar var med til at forstærke en storm af had mod denne befolkningsgruppe. Opildnet af opslag på Facebook, blev rohingyaerne dræbt, tortureret, voldtaget og fordrevet i tusindvis som en del af Myanmars sikkerhedsstyrkers kampagne for etnisk udrensning (Amnesty International, 2022; Stevenson, 2018).

De sociale medier er dog heller ikke nødvendigvis altid selv interesserede i det vilde vesten. Om end platformene generelt har haft en meget liberal linje ift. ekstremt indhold på deres platforme, heriblandt opfordringer til vold, så er faren for regulering, besværet med at stå skoleret for lovgivere og frygten for flugt af annoncører nogle af de incitamenter platformene har for at udføre indholdsmoderation. For eksempel oplevede Twitter (nu X) et tab på over 40 procent af deres annonceindtægter i månederne efter Elon Musks kontroversielle overtagelse og fremkomsten af et større antal antisemitiske tweets på platformen (Weiss, 2024). Et klart vink med en vognstang om, at annoncørerne, der betaler gildet, ikke vil finde sig i hvad som helst.

Selvreguleringen får dog ikke kritikerne til at sætte sig ned. Det er der primært to årsager til. På den ene side findes kritikken af *for meget* indholdsmoderation. Nogle oplever deres politiske holdninger censureret. Andre oplever, at de er underlagt amerikansk kulturimperialisme, hvor kvindelige brystvorter for enhver pris ikke må være synlige. På den anden side kritiseres sociale medier for *for lidt* indholdsmoderation. Russiske og kinesiske påvirkningskampagner og målrettet misinformation, vold og grænseoverskridende materiale, som rammer mindreårige (Børns Vilkår, 2024), samt hadtale mod etniske mindretal og kvinder (Zuleta & Burkal, 2017) er nogle af de emner, der har fået opmærksomhed i Danmark.

Forordningen om digitale tjenester (DSA) giver myndigheder, borgere og forskere en række værktøjer, når det kommer til at håndtere ulovlige og skadende aktiviteter samt arbejde for et mere fair og åbent onlinemiljø. EU har allerede gjort brug af lovgivningen og indledt undersøgelser af bl.a., hvorvidt X gør tilstrækkeligt for at undgå ulovligt indhold på deres

platform. Lever platformen ikke op til lovgivningen, kan den blive idømt en bøde på op til 6 procent af dens globale omsætning og i sidste instans blive forment adgang til det europæiske marked.

De nye regler, deriblandt kravet om at give bl.a. forskere adgang til data, går imidlertid fortsat træg. Der er kun få succeshistorier og det er stadig uklart, hvor præcist og effektivt et værktøj DSA bliver til at give dataadgang. Der er derfor behov for, at myndigheder, forskere og platforme finder ud af, hvad reglerne præcis indebærer, og hvordan de håndhæves. Men med DSA kan fremtidige undersøgelser forhåbentlig nå dybere ned i beskrivelser af, hvad der faktisk indholdsmodereres på platformene.

Forordningen om digitale tjenester (DSA) indebærer en lang række af nye regler. I hovedtræk indebærer forordningen:

- Klarere regler for fjernelse af ulovligt indhold
- Rettigheder til at klage over indholdsmoderationsafgørelser
- Mere transparens om anbefalingssystemer og annoncer
- Indskrænkning af hvem der kan målrettes annoncer
- Krav om årlige afrapporteringer af bl.a. indholdsmoderationsindsatser
- Krav om at dele data med bl.a. forskere
- Bedre sanktionsmuligheder for EU og medlemsstaterne

Aktive brugere på sociale medier og sociale mediers annonceindtægter

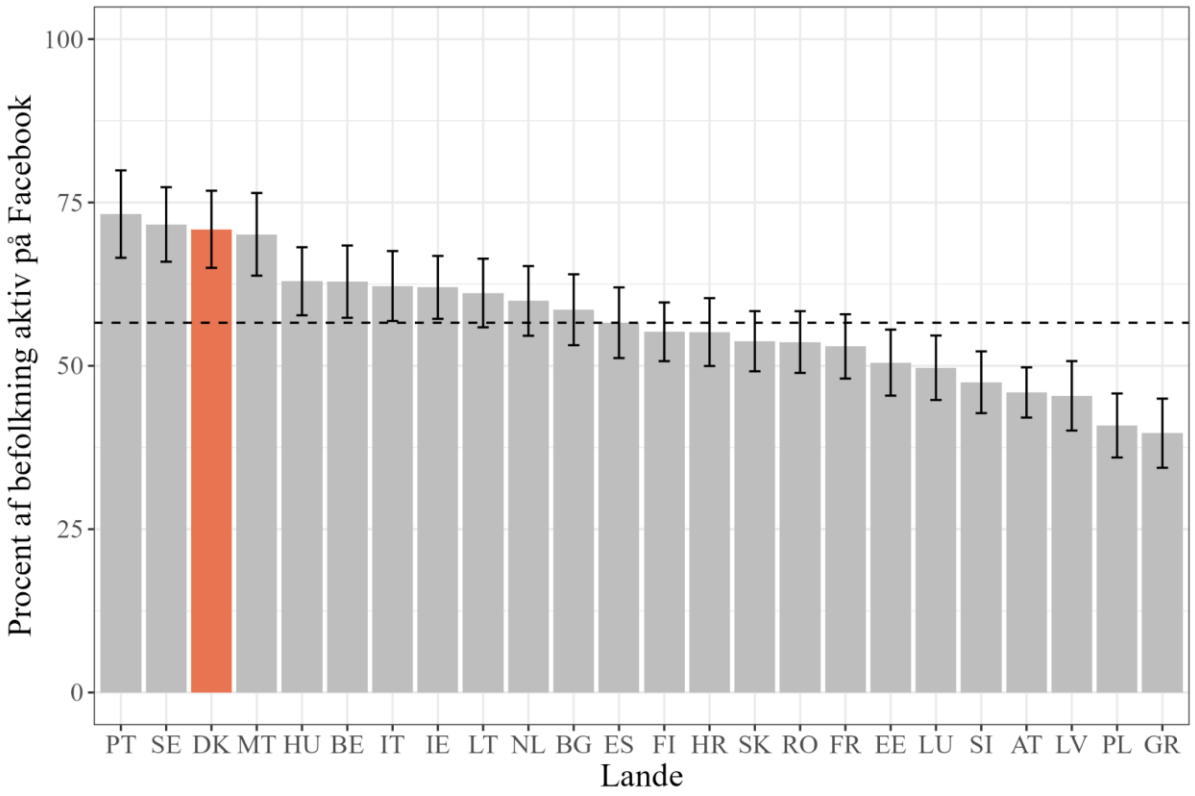
De sociale medier er meget tilbageholdende med at dele information om, hvem der er aktive på deres platforme, og hvor stor annonceomsætningen er i de enkelte lande. Den største udfordring, når det gælder sådanne opgørelser, er således datatilgængelighed, hvor ingen opgørelser er i stand til perfekt at sætte et tal på. Det bliver imidlertid lettere i de kommende år at estimere alle de større platformes annonceomsætning, da den europæiske forordning om digitale tjenester (DSA) indebærer et krav om oprettelse af såkaldte annoncebiblioteker, hvor information om annoncer på platformene kan findes.

Ud over den generelle udfordring med dataadgang, er der ligeledes stor variation mellem platforme i forhold til, hvilke data det er muligt at tilgå. Hvor Twitter i mange år var platformen med de bedst tilgængelige data for forskere og civilsamfundet, er Meta (selskabet bag bl.a. Facebook og Instagram) – lidt overraskende – blevet den platform med størst åbenhed om data. Vi fokuserer derfor i afsnittet her på Facebook. Ikke fordi det er ideelt, men fordi det er, hvad der er tilgængeligt på nuværende tidspunkt. Ligeledes kan tilgangen her anvendes fremadrettet til at undersøge, hvem der er aktive på platformene, og hvor meget de omsætter for på annoncer.

Data bag figurerne kommer fra Metas Marketing API og Metas annoncebibliotek API. Begge datakilder stiller data til rådighed for godkendte udviklere. Hvor man i førstnævnte datakilde kan dykke ned i, hvem brugerne på Metas platforme er, kan man i sidstnævnte datakilde se, hvilke annoncer der er at finde på Metas platforme.

Figur 1 viser en opgørelse over, hvor mange brugere[1] der er aktive på Facebook dagligt i alle EU-lande. For at kunne sammenligne mellem EU-landene er tallene sammenholdt med landenes befolkningstal. Markeret med orange er det muligt at se, hvordan Danmark ligger i toppen af de europæiske lande ift., hvem der har flest aktive brugere på Facebook. Specifikt er cirka 75 procent af den danske befolkning dagligt logget ind på Facebook. For de 13+ årige, som er aldersgrænsen for at oprette en bruger på Facebook, er procentdelen 81 procent. Ligeledes interessant er det, at der er stor variation i brugen af Facebook på tværs af Europa, hvor blot 40 procent af grækerne og polakkerne er logget ind på Facebook dagligt.

Figur 1: Procent af befolkningen, der dagligt er aktive på Facebook

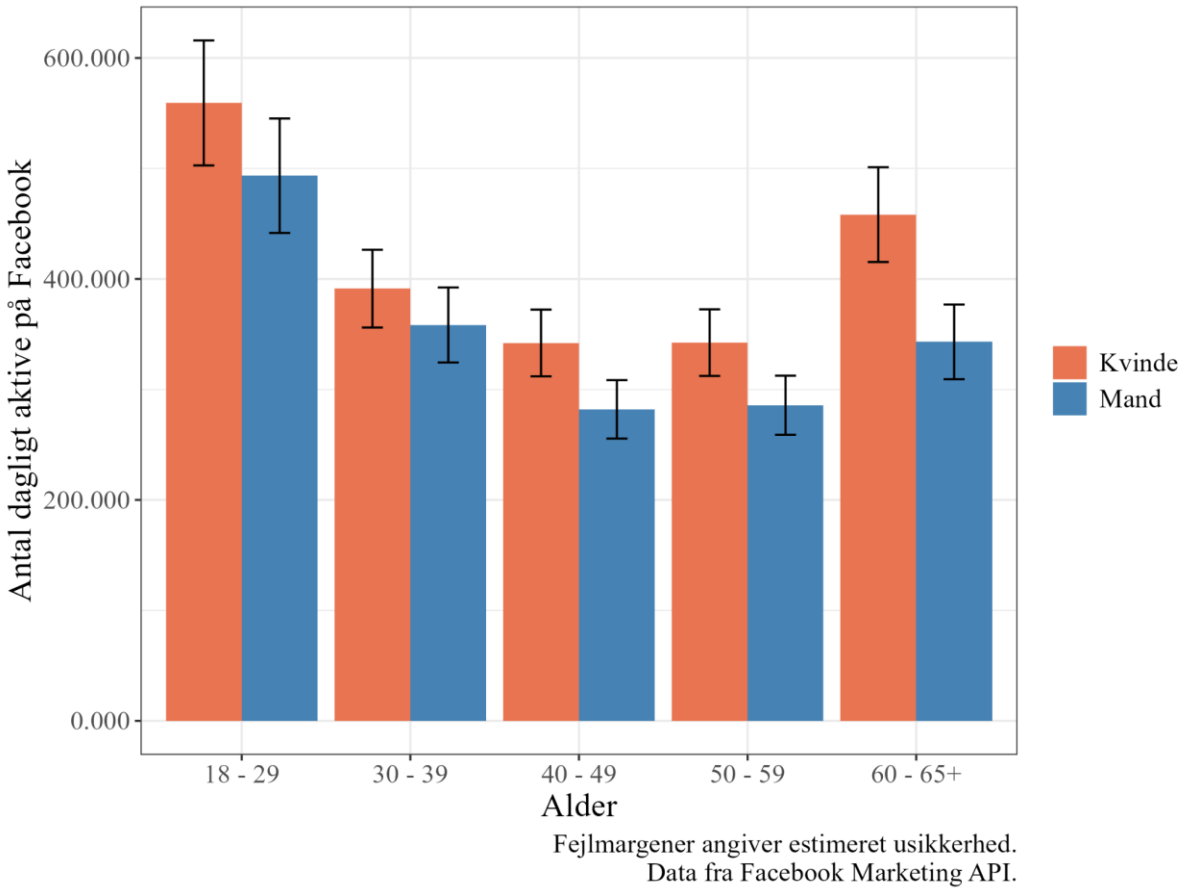


Fejlmargener angiver usikkerhed i estimatet.
Stiplet linje angiver gennemsnittet på tværs af EU.
Data fra Facebook Marketing API.

Figur 2 zoomer ind på hvilke befolkningsgrupper, der i højeste grad er til stede på Facebook, fordelt på alder samt mænd og kvinder. Som figuren viser, er der en overvægt af aktive Facebook brugere blandt unge samt ældre kvinder. Derudover er der generelt en overvægt af kvinder, der dagligt bruger Facebook i forhold til mænd.

[1] Meta oplyser antallet af aktive brugere og ikke personer, hvorfor tallet kan ses som en øvre grænse, da det er muligt for en person at have mere end én bruger. Ligeledes skal aktivitet forstås i en minimal forstand, således at en bruger vurderes aktiv alene ved at have åbnet f.eks. Facebooks app.

Figur 2: Unge, ældre og kvinder er oftere aktive på Facebook end midaldrende og mænd



Hvad angår annonceomsætning, har det længe været svært direkte at observere, hvor meget sociale medier omsætter for i de enkelte lande. Generelt har de, der har forsøgt sig med at estimere annonceomsætningen peget på, at en stigende del af annonceomsætningen for det første rykker online og for det andet i overvejende grad rykker mod tech-giganter som Google, Facebook og LinkedIn (Kulturministeriet, 2021). Samtidig viser opgørelser fra Danske Medier Research, hvordan annonceindtægterne hos dagbladene har været stødt faldende de senere år (Danske Medier Research, 2024).

Med EU-forordningen om digitale tjenester er tech-giganterne blevet pålagt at oprette såkaldte annoncebiblioteker, hvori alle annoncer på platformene skal offentliggøres. Mens bibliotekerne fortsat er i indledende stadier og at adgangen til deres API'er går trægt (Darius, 2024), er det ikke muligt systematisk at undersøge udviklingen i annonceomsætning på tværs af platforme i denne rapport. Fremtidige rapporter vil imidlertid forventeligt kunne bruge annoncebibliotekerne og fremgangsmåden fra rapporten her[2] til at estimere platformenes annonceomsætning på en relativ præcis og systematisk måde.

Vi fokuserer her på det, der er muligt direkte at observere, navnligt det udsnit af Metas annonceomsætning, der handler om samfundsmæssige forhold, hvor der i længere tid har været åbenhed[3] om annonceomsætning (se bl.a. Hove et al., 2024). Sådanne annoncer

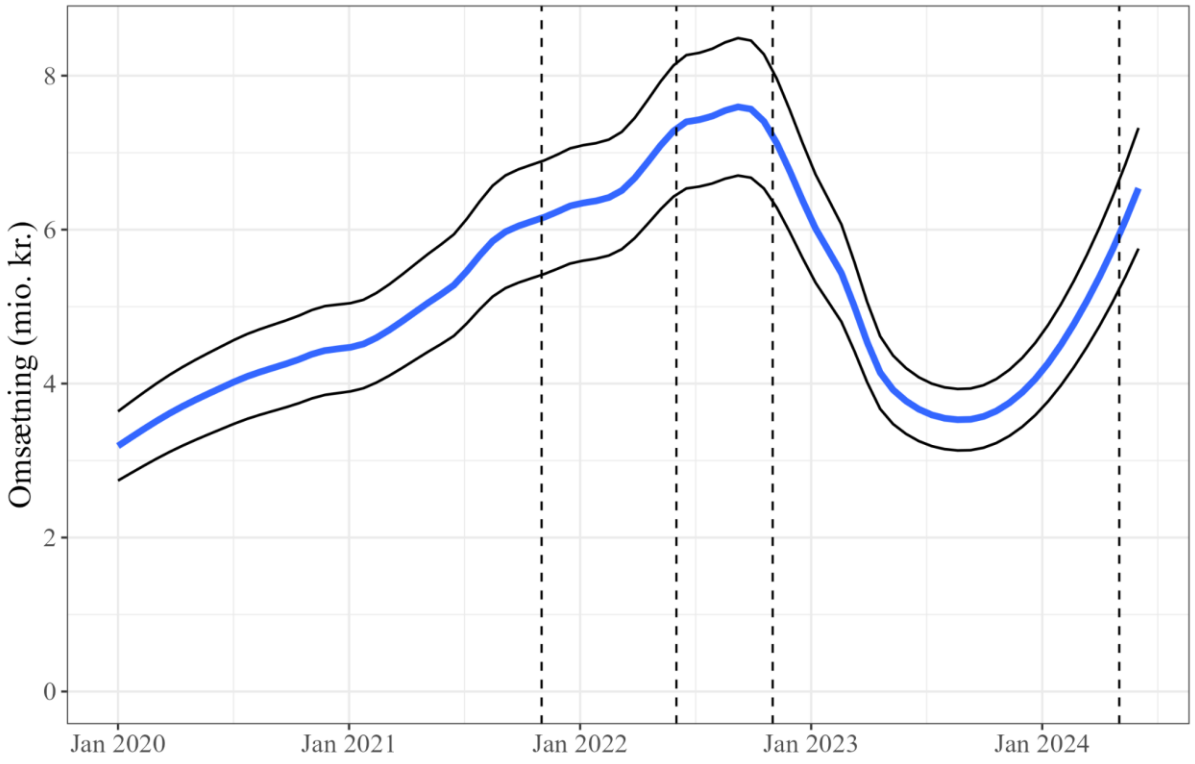
[2] Se bilag 1A for fremgangsmåde.
[3] For refleksioner om annoncebibliotekernes datakvalitet se bl.a.Leerssen (2023) og Dommett (2023)

handler om politik i bred forstand, f.eks. hvor annoncer, der advokerer for et bestemt politisk synpunkt eller omtaler et parti, også er inkluderet (Meta, 2024).

Annoncer om samfundsmæssige forhold på sociale medier har været til debat i mange år, idet der er frygt for, at de kan manipulere vælgere og påvirke demokratiske valg. Mens forskningen nedtoner den frygt af flere årsager (se bl.a. Hove, 2024) kan et overblik over, hvor stor aktiviteten er og hvor mange penge, der bliver brugt på annoncer, være med til at vise hvornår annoncerne indrykkes og for hvor mange penge annoncer relateret til den offentlige samfundsmæssige debat udgør af annoncemarkedet.

Figur 3 viser udviklingen i[4], hvor mange millioner kroner, der er blevet brugt i Danmark på annoncer om samfundsmæssige forhold på Facebook og Instagram i perioden januar 2020 til juni 2024. Figuren viser tydeligt, hvordan niveauet fra slutningen af 2021 til slutningen af 2022 har været højt sammenlignet med 2020 og 2023, samt at niveauet igen steg i foråret 2024. Udsvingene korrelerer i meget høj grad med, hvornår der har været valg, hvilket i figuren er markeret med stiplede linjer[5]. Det er derfor tydeligt, at størstedelen af annoncer om samfundsforhold indrykkes når danskerne skal til stemmeurnerne, om end der stadig er for mellem tre og fem millioner kroners omsætning uden for valgsæsoner.

Figur 3: Annonceforbrug om samfundsforhold er størst ved valg



Fuldtoptrukne sorte linjer angiver usikkerhed i estimat.
Vertikale stiplede linjer angiver tidspunkt for valg.
Data fra Meta annoncebibliotek API

[4] Udviklingen er udjævnet med lokal regression for at undgå store sæsonudsving ved f.eks. sommerferie.
[5] Kommunal- og regionsrådsvalg november 2021, afstemning om afskaffelse af forsvarsforbeholdet juni 2022, folketingsvalg november 2022 og europaparlamentsvalg juni 2024.

Figur 3 viser derudover en anden interessant pointe, nemlig at størrelsesordenen for samfundsmæssige annoncer er meget lav sammenlignet med estimater for Facebooks samlede annonceomsætning. Således når de samfundsmæssige annoncer for perioden januar 2020 til juni 2024 en omsætning på knap 300 mio. kr., mens estimatet for Facebooks samlede annonceomsætning i Danmark alene i 2020 er 1.387 millioner kroner (Kulturministeriet, 2021). Annoncer om samfundsmæssige forhold er derfor ikke alene meget svingende, de udgør også en lille del af det samlede annoncemarked.

Vurderinger

At komme med vurderinger, når det gælder tech-giganters indflydelse på samfundet, kræver at man forholder sig aktivt til en definition af demokrati. Hvad nogen anser som sundt og for demokratiet nødvendig indholdsmoderation, anser andre for censur. Vores analyser og vurderinger beror på en forståelse af demokrati, som omfatter at alle har adgang til retvisende information om nyheder og politik, og hvor det er muligt at deltage i offentlige samtaler uden at frygte trusler og diskrimination.

Vi har følgende vurderinger på baggrund af rapportens første del.

Det er vigtigt med et pluralistisk sæt af udbydere af digitale services, inklusive medier. Et divers økosystem er til gavn for både markedsøkonomi og demokrati, for bl.a. at kunne sikre, at medierne kan udfylde deres rolle som offentlighedens vagthund. Dette er ligeledes i tråd med den nyligt vedtagne europæisk mediefrihedslov (EMFA), der pålægger medlemsstaterne at sikre et divers mediemiljø og uafhængig journalistik.

Sikre arbejdsgange, så eksisterende lovgivning kan håndhæves og anvendes effektivt af myndigheder, forskere og civilsamfund. Lovgivningen i bl.a. DSA og DMA er afgørende for at kunne undersøge samfundsmæssige betydninger af tech-giganter. Men det kræver politisk vedholdenhed, hvis lovgivningens potentiale skal indløses til gavn for myndigheder, forskere og civilsamfund. Det kræver bl.a. investeringer i effektiv information om lovgivning og ressourcer til håndhævelse af de nye regler samt hjælp til forskere, myndigheder og andre til at skaffe adgang til den data, der muliggøres i lovgivningen.

Manglende dataadgang er den store hæmsko, særligt hvad angår viden om danske forhold. I forlængelse af vurderingen om behovet for at sikre, at eksisterende lovgivning udnyttes, er det afgørende at fokusere på, at de lovgivninger og programmer, der er sat i søen, ikke alene er et fælleseuropæisk ansvar, men også et dansk. Den næste fase er vigtig for at sætte indsatser og programmer i gang, der skal sikre, at den viden, der bliver produceret, også tager udgangspunkt i Danmark. Det vil styrke vores evidensbase for at kunne diskutere, hvordan tech-giganter påvirker det danske samfund, og hvilke passende initiativer der kan adressere de eventuelle udfordringer.

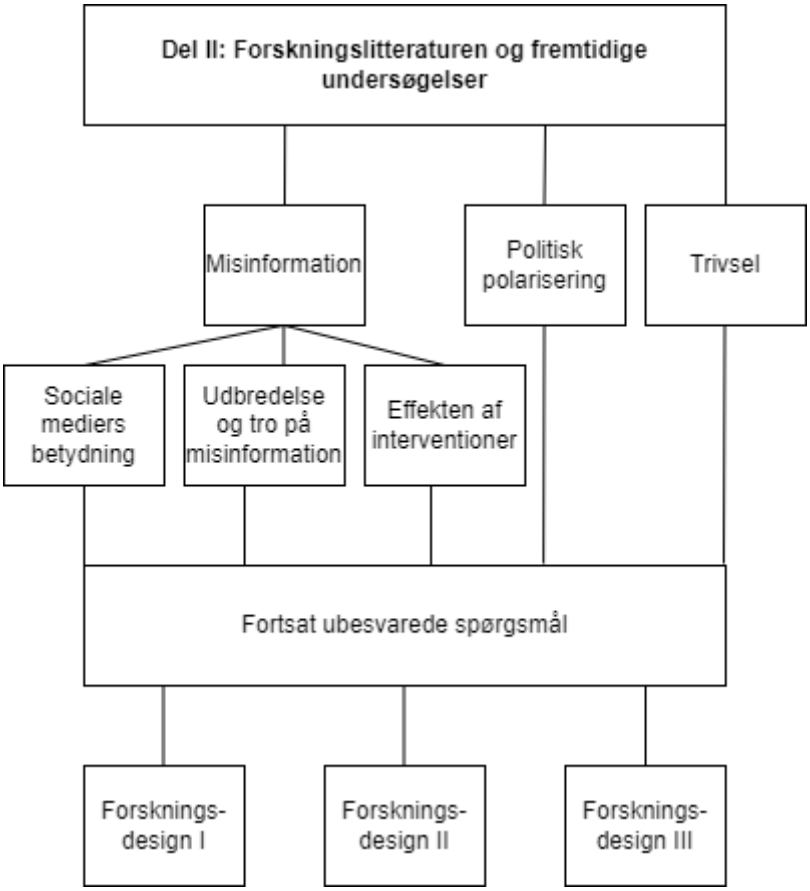
Hvad ved vi om sociale mediers betydning for misinformation, politisk polarisering og trivsel?

At give et overblik over, hvad vi fra forskningen ved og ikke ved om, hvordan sociale medier påvirker vores demokrati, vores velbefindende eller sammenhængskraft kræver, at man holder tungen ualmindeligt lige i munden. Selvom dygtige forskere verden over bruger dag ud og dag ind på at undersøge betydningen af sociale medier, så er opgaven nærmest uoverstigelig. Det har tidligere været – og er stadig – svært at få adgang til de nødvendige data, deriblandt fra platformene selv. Forskere er derfor oftest overladt til spørgeskemaer, observationer og andre metoder, der kan give indikationer, men som har svært ved at lade sig oversætte direkte til den virkelighed, som udspiller sig. Problemet er i den sammenhæng, at vi ikke har nogen verden, der løber parallelt med vores, hvor sociale medier ikke findes, som vi kan sammenligne med. At tro, man uden videre kan skære sociale medier ud af ligningen og isolere effekten, er naivt.

Det er derfor let at komme til at fejltolke (ellers gode og veldokumenterede) forskningsresultater. Hvis en gruppe af mennesker logger af sociale medier i tre uger og derefter er i bedre humør end dem der ikke loggede ud, så betyder det ikke, at det er de sociale medier, der gør folk mere kede af det. Men det betyder heller ikke, at de ikke gør det. De viser kun, at i en verden, hvor vi bruger sociale medier hver dag, og hvor resten af vores venner og bekendte forbliver online, selvom vi logger af i tre uger, så får vi det bedre. Vi har ikke en parallel verden uden sociale medier, vi kan sammenligne med.

Formålet med denne del af rapporten er at danne et overblik over, hvad vi fra forskningen ved om sociale mediers betydning for misinformation, politisk polarisering og trivsel. Med den viden i hånden kan vi efterfølgende udpege vigtige spørgsmål, vi fortsat mangler viden om, samt hvordan de vil kunne besvares.

Forskningslitteraturen er stor og indsigterne er mange. Som følge af rapportens primære fokus på misinformation fylder denne del mest med flere underpunkter. Derudover, i tråd med Medieaftalens fokus, er der to kortere afsnit om emnerne politisk polarisering og sociale mediers betydning for trivsel, der ligeledes er oplagte fokusområder for fremtidige rapporter. Alle tre dele opsamles i en række fortsat ubesvarede spørgsmål på bagkant af litteraturgennemgangen, der ligeledes munder ud i tre konkrete forskningsdesigns, der kan tages op i fremtidige rapporter. Strukturen for forskningsgennemgangen og de efterfølgende forskningsspørgsmål er opsummeret i figuren nedenfor mens metoden til at indsamle og systematisere litteraturen er beskrevet i bilag 2A.



Misinformation på sociale medier

At nogle spreder misinformation er ikke noget nyt fænomen. Allerede i 1600-tallets Frankrig kunne borgerne på gaderne i hovedstaden købe en såkaldt "canard". En avis, der mættede befolkningens sult efter sensationelle og underholdende historier – der til gengæld ofte var falske (Baptista & Gradim, 2020).

I dag er misinformation et koncept, som de fleste af os kender. Godt hjulpet på vej af den tidligere amerikanske præsident, Donald Trump, der hyppigt omtalte nyhedsmedier, han ikke var enig med som *fake news*. Eller russiske troldes forsøg på at påvirke valget med falske historier med alt fra at pave Frans opfordrede amerikanere til at stemme på Trump, til at Hillary Clinton havde solgt våben til Islamisk Stat.

Frygten fra eksperter og befolkninger verden over er, at misinformation slæber dårligdomme med sig: påvirkning af demokratiske valg, svækket sammenhængskraft og lavere tillid til medier og politikere. Ved indgangen til det store valgår 2024 var hele 87% af befolkningerne i de lande, der skulle stemme til parlamentsvalg, nervøse for, at misinformation vil påvirke valget (Quétier-Parent et al., 2023). Og World Economic Forum udpegede på baggrund af en spørgeskemaundersøgelse blandt eksperter desinformation som den største udfordring for samfundet på kort sigt (World Economic Forum, 2024).

Kært barn har som bekendt mange navne, og forkerte informationer er ingen undtagelse. Hyppigt hører man dem omtalt som *fake news*, misinformation og desinformation. Om end

ordene dækker over vigtige nuancer – f.eks. hvorvidt informationen med vilje er falsk – holder vi os for overskuelighedens skyld til her at alene omtale misinformation, bredt defineret som falske påstande, der fremlægges som korrekte (Allcott & Gentzkow, 2017: 213)**.**

Sociale mediers betydning for spredning af misinformation

Linjen mellem professionel og amatør sløres yderligere på sociale medier

Det indhold, man som bruger på sociale medier møder, er skabt af en mere divers skare af såkaldte "indholdsskabere", end man er vant til hos mere traditionelle medier. Dette gør kun linjen mellem, hvem der er professionel og hvem der er amatør, mere uklar. Kombineret med mulighederne for ofte at være anonym, inviterer sociale medier derfor aktører inden for, som kan være interesserede i at dele misinformation, og som ville have haft sværere ved at udkomme i et ikke digitalt format (Kim et al., 2021; Shahzad et al., 2021).

Algoritmisk kuration af nyheder risikerer at øge spredningen af misinformation

Det er imidlertid ikke alene, hvem der producerer indhold på sociale medier, der øger risikoen for, at misinformation spredes. Sociale medier afviger væsentligt fra traditionelle medier, når det gælder, hvad man som bruger bliver eksponeret for. Sociale medier er opbygget med algoritmiske *feeds*, som påvirker, hvad der bliver vist til den enkelte bruger. På den måde er der en risiko for, at brugernes efterspørgsel på og tiltrækning til sensationelle nyheder baner vejen for spredning af misinformation (Akram et al., 2022; Shahzad et al., 2021). Dette er ligeledes konklusionen på et af de klassiske (og omdiskuterede) studier inden for feltet, hvor forskere finder, at verificerede falske nyheder breder sig betydeligt længere, hurtigere, dybere og mere bredt end verificerede sande nyheder på Twitter (Vosoughi et al., 2018).

Algoritmiske *feeds* er en måde, hvorpå brugerne vises indhold på baggrund af, hvad de tidligere har klikket på, hvem de er venner med og hvad andre klikker på. På den måde har opslag og billeder, der oplever stor tiltrækningskraft, mulighed for at gå viralt, hvor millioner af mennesker ind i mellem bliver vist det samme opslag.

I hvor høj grad algoritmiske *feeds* påvirker, hvad brugerne ser, og hvordan det påvirker deres holdninger, er imidlertid endnu ikke afgjort. Der er flere, der udfordrer idéen om, at algoritmerne er drivende for at misinformation bliver spredt og set på sociale medier. For eksempel viser et amerikansk studie, at de amerikanere, som ser indhold fra ekstremistiske kanaler på YouTube, i forvejen har sexistiske holdninger og høj grad af racemæssig modvilje. Ligeledes observerer forskerne, at brugere der bliver blokeret for at interagere med for meget misinformation blot rykker over til andre platforme for at opsøge samme indhold der (Budak et al., 2024). Lignende eksempler ser vi også i Danmark, hvor et dansk studie blandt andet viser, at folk, der er fjendtlige på sociale medier, også er det i virkelighedens verden (Bor & Petersen, 2022). Det er derfor svært direkte at adskille, hvad der i spredningen af misinformation skyldes de sociale mediers algoritmer, og hvad der skyldes brugernes efterspørgsel.

Sociale mediers forretningsmodeller deincentiverer kampen mod misinformation

En anden central forklaring på, hvorfor misinformation kan få lov til at blive spredt på sociale medier, handler om sociale mediers forretningsmodeller. Tech-giganterne bruger i varierende

omfang den data, de indsamler om brugerne, til at optimere deres egen forretning samt til at videresælge bl.a. til virksomheder, der ønsker at reklamere til bestemte segmenter (Aral, 2021). Tiltrækningskraften ved misinformation risikerer således at ende med at være et gode, der skaber kliks og opmærksomhed, og i sidste ende større omsætning (Kaushik, 2024; Sanders & Jones, 2018). Incitamentet for platformene står derfor ikke soleklart, idet det både er besværligt og potentielt dyrt at mindske mængden af misinformation på platformene.

Udbredelse og tro på misinformation

Økonomi, ideologi og underholdning driver de, der skaber misinformation

De, der skaber misinformationsindhold, er ofte drevet af økonomiske, ideologiske og underholdningsmæssige motiver (Wu et al., 2024; Kim et al., 2021; Baptista & Gradim, 2020). Det økonomiske motiv handler oftest om at lokke folk ind på en hjemmeside ved at skabe sensationelle *clickbait* overskrifter, hvor annoncevisninger genererer indkomst til misinformationsaktørerne. Et andet eksempel var tidligere i år højt på mediedagsordenen i Danmark, hvor falske kendis-annoncer på Facebook bl.a. forsøger at svindle folk med lovning om hurtige økonomiske gevinster (Nisgaard, 2024). Og det er ikke helt uden held. NewsGuard, der bl.a. vurderer troværdigheden af hjemmesider, estimerer at det økonomiske marked for misinformation løber op i ca. 18 milliarder kroner årligt alene i annonceindtægter (Skibinski, 2021). Godt hjulpet på vej af almindelige virksomheder, der uvidende annoncerer på hjemmesider, der er kendt for at sprede misinformation (Ahmad et al., 2024).

Det ideologiske motiv er derimod det, de fleste af os nok vil forbinde med misinformation. Her spredes misinformation med henblik på at gavne foretrukne politikere og tilsmudse politiske modstandere. Det kan både være falske nyhedssider som Breitbart News, der forsøger at hjælpe republikanerne ved at sprede falske nyheder om demokraterne, ligesom det kan være fremmede stater som Rusland, der forsøger at påvirke stemningen og valgresultater i deres foretrukne retning (Golovchenko et al., 2020).

Til sidst er der dem, som spreder misinformation og *troller* alene med henblik på deres egen og andres underholdning.

Misinformation udgør en lille andel af nyhederne, men er også svært at måle

I forskningslitteraturen er der ikke fuldkommen enighed om, hvor stor en udfordring misinformation er. Når det forsøges at undersøge, hvor meget misinformation fylder på sociale medier, er en af udfordringerne, at det ofte er svært at nå til enighed om, hvordan vi forstår misinformation og at måle, hvad der er misinformation eller hvad der ikke er. Afhængigt af hvordan misinformation defineres og måles, er der derfor både risiko for at underestimere den sande mængde af misinformation (Pennycook & Rand, 2021) og overestimere den (Budak et al., 2024). En anden udfordring er, at vores viden er begrænset af, at det meste forskning fokuserer på USA, der med meget høj grad af polarisering er en særlig *case*. Det kræver derfor, at man er påpasselig når man forsøger at oversætte amerikanske resultater til Danmark.

Tidligere forskning estimerer, at falske nyhedssider i gennemsnit står for mellem 0,7% og 6% af de nyhedshistorier, der *linkes* til på sociale medier (Altay et al., 2022). For at sætte i relation til absolutte tal, så var der, i de tre måneder op til det amerikanske præsidentvalg i 2016,

mindst 38 millioner delinger på Facebook af falske nyheder (Allcott & Gentzkow, 2017) ligesom opslag fra russiske internettrolde havde en rækkevidde på op til 126 millioner amerikanske statsborgere på Facebook (Budak et al., 2024) og sendte 109.000 tweets (Golovchenko et al., 2020).

Hvorvidt de tal er store modtager i forskningslitteraturen typisk to typer af indvendinger. For det første er der betydningen af selektion. Typen af folk, der deler og interagerer med misinformation er ofte en afgrænset gruppe (Budak et al., 2024). Ny forskning viser bl.a., hvordan 80 procent af alle delinger af falske nyheder på Twitter under det amerikanske præsidentvalg i 2020 blev delt af 2.107 registrerede vælgere, svarende til 3 promille af amerikanske Twitter-brugere (Baribi-Bartov et al., 2024). Ligeledes udgør falske nyheder kun en lille del af, hvad den gennemsnitlige person på sociale medier "synes-godt-om", deler eller klikker på (Pennycook & Rand, 2021, Baptista & Gradim, 2020: 11).

Den anden indvending tager udgangspunkt i, at folk generelt bliver eksponeret for rigtig meget information på internettet. Tager man de 126 millioner amerikanere, der muligvis[6] har set opslag fra russiske internettrolde, så udgør det kun 0,004% af det indhold, der blev vist i deres Facebook *feed* (Budak et al., 2024). Tager man estimater af al online misinformation, og ikke kun det på Facebook eller det af russiske trolde, er estimatet, at misinformation udgør 0,15% af hvad amerikanere og 0,16% af hvad franskmænd ser online (Altay et al., 2023).

Der er derfor meget der tyder på, at eksponeringen for misinformation udgør en lille andel af den information, folk modtager på sociale medier, og at den særligt er koncentreret hos en mindre gruppe.

Om end det kan lyde positivt, er det ikke ensbetydende med, at man kan forklejne udbredelsen af misinformation. For det første har misinformationen det særlig godt, når der er kriser eller meget usikkerhed, hvor det er sværere at gennemgå og tjekke informationen (Akram et al., 2022). For det andet spredes misinformation også fra ellers tillidsvækkende aktører og på en måde, hvor misinformationen er subtil. En nylig forskningsartikel udlægger tydeligt den udfordring. Konkret viser forskerne, hvordan identificeret misinformation om Covid-19 fik 8,7 millioner visninger på Facebook i de første tre måneder af 2021, mens faktuelt korrekte, men implicit vildledende vaccineskepsis blev vist flere hundrede millioner gange med en markant større overbevisningsevne (Allen et al., 2024). For det tredje viser en anden ny forskningsartikel, at russisk propaganda om geopolitik jævnligt spredes af en bred skare af amerikanske medier på begge ideologiske fløje (Yang et al., 2024). Det er derfor vigtigt at huske, at blot fordi noget er svært at måle, så betyder det ikke, at effekten er lig nul.

Uopmærksomhed og mentale genveje påvirker om folk tror på misinformation

Når det kommer til, hvorfor folk tror og accepterer misinformation, peger forskningen typisk på to forklaringer. Den første forklaring går på, at folk deler misinformation, fordi det stemmer overens med deres politiske verdenssyn, mens den anden peger på, at folk deler misinformation, fordi de ikke er opmærksomme nok, når de ser og deler information på sociale medier. Om end forskningen peger på, at begge faktorer er i spil, så tyder meget på, at

[6] Det er ikke muligt at afgøre, hvorvidt en person faktisk har set opslaget, men alene, at det har været præsenteret på personens *feed*.

spørgsmålet om folks opmærksomhed er den stærkeste af de to faktorer (Pennycook & Rand, 2021; Kim et al., 2021).

Derudover peger forskningen på, at når det går stærkt med informationen på sociale medier findes måden, hvorpå folk falder for misinformation, ofte i hvad kaldes *heuristikker* – mentale genveje hjernen bruger til hurtigere at nå frem til en konklusion. Heriblandt nævnes særligt tre af sådanne genveje. For det første er man mere tilbøjelig til at tro på (mis)information, hvis den er genkendelig (Wu et al., 2022; Pennycook & Rand, 2021; Greenspan & Loftus, 2020). Har man eksempelvis tidligere hørt et rygte, er man også mere tilbøjelig til at vurdere det som sandfærdigt. For det andet reagerer folk ofte på social *feedback*. Ser man, at et opslag på sociale medier har mange *likes* eller delinger, er der større sandsynlighed for, at man finder opslaget troværdigt, da mange andre mennesker tilsyneladende interagerer med indholdet (Wu et al., 2021; Pennycook & Rand, 2021; Bryanov & Vziatysheva, 2021). For det tredje indgår en persons følelser og personlighed i ligningen. Eksempelvis peger forskningen på, at folk, der afrapporterer en høj grad af positivitet eller negativitet, er mere tilbøjelige til at tro på falske nyheder (Wu et al., 2021; Pennycook & Rand, 2021; Bryanov & Vziatysheva, 2021).

Uklart, hvor gode folk er i stand til at skelne mellem sandt og falsk på sociale medier

Tæt relateret til spørgsmålet om, hvorvidt folk tror på misinformation er, hvorvidt de er i stand til at skelne, hvad der er sandt fra hvad der er falsk. Det er imidlertid svært på samme tid at indfange noget virkelighedsnært og undgå alene at teste folks evne til at huske, hvad de har hørt om i nyhederne. Derfor er resultaterne også blandede i forhold til den endelige konklusion om evnen til at kunne skille sandt fra falsk. Nogle studier finder, at folk ikke klarer sig bedre end tilfældighed, mens andre studier viser, at folk i høj grad er i stand til at skelne (Bryanov & Vziatyshea, 2021).

Årsagen er måske at finde i, om man er opmærksom og evner at tænke analytisk. Således er folk, der er mere analytisk tænkende og opmærksomme, i højere grad i stand til at adskille sand fra falsk information, ligesom de er bedre til at vurdere hvilke nyheder, der er politisk *biased,* end de mindre analytisk tænkende og opmærksomme (Pennycook & Rand, 2021; Bryanov & Vziatysheva, 2021; Baptista & Gradim, 2020). Derudover er folk bedre til at adskille sandt fra falsk når det gælder politik end når det gælder emner, hvor de fleste generelt har mindre viden, som f.eks. sundhed og forskning (Bryanov & Vziatysheva, 2021). Og selvom folk i højere grad stoler på information, der er politisk kongruent med deres holdninger, så er de også bedre til at identificere misinformation i politisk kongruent information. Overordnet indikerer dette, at troen på misinformation kommer fra uopmærksomhed, og ikke fordi de er blevet politisk *hijacked* til at tro på noget (Pennycook & Rand, 2021).

Udover opmærksomhed og evnen til at tænke analytisk, peger forskningslitteraturen også på andre forskelle. De, der er ringere til at skelne, har ofte ringere digitale færdigheder, kortere uddannelse, stærke ideologiske holdninger og mistro til medier, ligesom ældre og mænd oftere tror på misinformation (Baptista & Gradim, 2020).

Effekten af interventioner mod misinformation

Afsnittet her fokuserer på mulige interventioner, der kan bruges mod misinformation. Konkret kigger vi på, hvad forskningen siger om effekterne af forskellige interventioner. Vi behandler

derfor ikke den del af forskningen, der handler om, hvordan man kan identificere misinformation online.

Faktatjek er en effektiv intervention, men er ikke særlig skalerbar

En af de mest undersøgte metoder til at bekæmpe misinformation er faktatjek, som ofte markeres med advarselstegn. Advarselstegn kan inkludere en gul advarselstrekant eller et rødt kryds og en beskrivelse af, at professionelle faktatjekkere har vurderet indholdet falsk. Forskningsresultater viser, at sådanne interventioner kan reducere troen på, delinger af og interaktion med misinformation (Martel & Rand, 2023; Pennycook & Rand, 2021). Særligt tre ting påvirker, om og hvor godt interventionen virker. Advarselstegn er effektive, når de er placeret prominent, tydeligt markerer indhold som falsk og der indgår en forklaring på, hvad der er forkert i den information man ser (Martel & Rand, 2023; Morrow et al., 2021; Greenspan & Loftus, 2020; Walter & Tukachinsky, 2020). Effekten er desuden lille, når korrektionen kommer fra andre borgere, sammenlignet med når den kommer fra nyhedsorganisationer og eksperter (Walter et al., 2021; Walter & Tukachinsky, 2020). Samtidig afhænger effekten af emnet, der er tale om, hvor det er lettere at intervenere mod misinformation om kriminalitet og sundhed og i mindre grad om *marketing* og politik (Walter & Murphy, 2018).

Tidligere var der i forskningen bekymringer om, at faktatjek kunne have modsatrettede effekter, kendt som *backfiring,* hvor folk bliver endnu mere overbeviste om deres falske tro. Nyere undersøgelser indikerer dog, at sådan *backfire* sker meget sjældent (Morrow et al., 2022; Bryanov & Vziatysheva, 2021). Forskningen peger imidlertid på, at faktatjek ikke fjerner troen på misinformation fuldstændig, da ikke alle ser dem (Martel & Rand, 2023), folk har det med at huske (mis)informationen lettere end faktatjekket (Baptista & Gradim, 2020; Greenspan & Loftus, 2020), og advarselstegn har svært ved misinformation, som bliver gentaget, da det ikke er hver gang, der kommer et nyt advarselstegn på (Martel & Rand, 2023, Morrow et al., 2022).

Om end faktatjek kan være effektivt, er der flere udfordringer forbundet med metoden. Faktatjek er ikke let skalerbart, da det kræver personer, som sidder og tjekker indhold, hvilket gør det svært at udrulle bredt. En anden udfordring er risikoen for at skabe "implicit sandhed", hvor fraværet af advarselstegn kan få folk til at tro, at informationen er sand, selvom den ikke nødvendigvis er det (Martel & Rand, 2023). En tredje udfordring er, at effekten af faktatjek blandt andet afhænger af platformen og mediet, hvorfor nyere udfordringer som *deepfakes* og generativ AI risikerer at gøre faktatjek mere besværligt (Martel & Rand, 2023; Morrow et al., 2021).

'Vaccination' mod misinformation som innovativ intervention

En anden tilgang til at tackle misinformation, der særligt har vundet indpas i senere år, er *prebunking*, også kendt som *inoculation* teknikker. Disse teknikker handler om at lære folk at genkende de strategier, som ofte er til stede i misinformation, såsom polariserende sprog (Pennycook & Rand, 2021). Formålet er således at "vaccinere" befolkningen mod at falde for falske nyheder ved at øge deres kritiske tænkning og grad af *digital literacy*.

Forskning har vist, at sådanne teknikker kan forbedre folks evne til at identificere falske nyheder markant og at dette virker bedre end de faktatjek, der sker efter at en person har været eksponeret for misinformation (Greenspan & Loftus, 2020). For eksempel blev deltagere

i USA og Indien henholdsvis 26% og 19% bedre til at identificere falske nyheder efter at have modtaget træning i at genkende misinformation (Bryanov & Vziatysheva, 2021).

Selvom *inoculation* har vist sig effektivt, løser det ikke nødvendigvis det grundlæggende problem med, at folk ofte deler information uden at tage sig tid til at vurdere dens rigtighed. Derudover har løsningen ligeledes udfordringer med skalerbarhed, hvor det vil være besværligt at få folk verden over til at tage et sådan træningsforløb.

***Accuracy prompts* og *crowdsourcing* er mere skalerbare, men ikke uden udfordringer**

Af mere skalerbare tilgange til bekæmpelsen af misinformation findes særligt interventionerne *accuracy prompts* og *crowdsourcing*.

Accuracy prompts indebærer, at folk bliver bedt om at overveje, om den information, de er ved at dele, er sand, inden de får lov til at trykke send (Pennycook & Rand, 2021; Greenspan & Loftus, 2020). Metoden kan være effektiv, da den får folk til at stoppe op og reflektere over indholdets troværdighed, hvilket mindsker spredningen af misinformation (Morrow et al., 2021). Konkret sænker *accuracy prompts* intentioner om at dele falske nyheder med cirka 10 procent, ligesom de, der har modtaget en sådan *prompt,* er 72% bedre til at identificere falske nyheder end de, der ikke modtog en (Pennycook & Rand, 2022). Der er ingen forskel på effekten fra *accuracy prompts* på tværs af køn, etnicitet, ideologi, uddannelse eller ønske om præcision. Til gengæld er effekten større hos ældre, folk, der scorer højere på en kognitiv reflektionstest, og de som generelt er mere opmærksomme (Pennycook & Rand, 2022).

Crowdsourcing er, hvor almindelige brugere markerer indhold, de mener er misvisende (Pennycook & Rand, 2021). Denne måde at markere misinformation på er særlig tiltalende fordi den er væsentligt mere skalerbar og fortsat er effektiv til at sænke troen på misinformation, omend ikke i lige så høj grad som faktatjekket indhold (Martel & Rand, 2023). Selvom der kunne være en forventning om, at politiske skel ville påvirke vurderingerne, er folk generelt enige om, hvad der er lav- og højkvalitetsnyheder på tværs af politiske ståsteder (Pennycook & Rand, 2021). Dog kræver dette system bl.a., at brugerne er bekendt med hvilke medier, der er kendt for at sprede misinformation, hvilket ikke altid er tilfældet, især når det gælder mere niche-orienterede medier.

Metoderne er forbundet med en række udfordringer. For det første overkommes udfordringen med "implicit sandhed" ikke, da det stadig ikke realistisk vil være muligt at dække alt indhold på platformene. For det andet er der større risiko for *tainted truth*, hvor forkert-placerede advarsler på sand information reducerer tilliden til korrekt information generelt (Martel & Rand, 2023).

Effektiviteten af advarselstegn varierer også på tværs af platforme og typer af medier. For eksempel er *memes* og *deepfakes* særligt udfordrende at mærke korrekt. Forskellige typer advarselstegn har også varierende grader af succes, hvor detaljerede advarsler generelt er mere effektive end generelle advarsler. Derudover er det afgørende, at advarselssystemer er gennemsigtige og retfærdige for at undgå beskyldninger om censur og bevare offentlighedens tillid.

***Deplatforming* kan potentiel være effektivt, men er en balancegang med oplevet censur**
En anden ofte diskuteret mulighed, som primært ligger i hænderne på platformene selv, er

muligheden for *deplatforming*. Dette indebærer bl.a. muligheden for blokering og suspension af brugere, der spreder misinformation (Nasery et al., 2023). Et nyt studie undersøger betydningen af, at det sociale medie Twitter *deplatformede* et stort antal af brugere, der delte misinformation, inklusiv den daværende amerikanske præsident Donald Trump (McCabe et al., 2024). Studiet viser bl.a., hvordan interventionen mindskede mængden af misinformation, der cirkulerede på platformen, ligesom en række andre brugere med hang til at sprede misinformation, men som ikke var blevet *deplatformed,* forlod Twitter kort tid efter.

Der er imidlertid fortsat få empiriske studier og få stærke kausale estimater (Golovchenko, 2022; King et al., 2013). Samtidig er der anden forskning, der viser, at selvom Facebook fjernede anti-vaccine indhold under Covid-19 pandemien, så medførte det ikke en reduktion i antallet af interaktioner med anti-vaccine indhold (Broniatowski et al., 2023). Hvad angår *deplatforming* som strategi er det derfor fortsat uafklaret om det er en effektfuld strategi.

En udfordring ved *deplatforming,* såvel som anden indholdsmoderation, er risikoen for, at brugerne oplever det som censur (Nasery et al., 2023). Generelt giver brugere af sociale medier udtryk for både et ønske om, at platformene skal gøre mere for at fjerne misinformation, og samtidig ønsker mindre censur (Morrow et al., 2021). Det samme mønster ser vi gældende i Danmark, hvor danskerne er overvejende enige om behovet for indholdsmoderation på online platforme, men mere splittede på spørgsmål om, hvorvidt indholdsmoderation er skadeligt for ytringsfriheden (Center for Sociale Medier, Tech og Demokrati, 2023).

Hvor vi nu har gennemgået centrale dele af litteraturen om misinformation, skifter vi spor til to kortere afsnit, der redegør for de overordnede fund og diskussioner indenfor forskningslitteraturene om henholdsvis sociale mediers betydning for trivsel og politisk polarisering.

Politisk polarisering

Et af de mest omdiskuterede emner, hvad angår effekterne fra sociale medier, handler om politisk polarisering. Hovedargumentet er, at sociale medier og deres algoritmer faciliterer ekkokamre, som ændrer på hvilken information en person bliver eksponeret for og således splitter både politikere og befolkningen i ideologiske grupper (Iyengar et al., 2019).

Forskningen har imidlertid været splittet om, hvor stor en skyld sociale medier bærer for politisk polarisering. Forskning, der undersøger folks holdninger ved hjælp af spørgeskema, finder en sammenhæng mellem at have polariserede holdninger og det at modtage nyheder fra medier, man politisk er enig med (Garrett et al., 2014, Lu & Lee, 2019). Ligeledes viser flere eksperimentelle studier, hvordan for eksempel eksponering for informationskilder, man ideologisk er på linje med, kan have en polariserende effekt (Levendusky, 2013).

Udfordringen er imidlertid, at det at stille folk spørgsmål i et spørgeskema eller teste kort og simuleret indhold i en kunstig situation, ikke kommer ind til kernen af problemet. Det er for eksempel svært at vide, hvor nemt sådanne fund generaliseres og hvor lang tid effekterne holder ved. Skal man ind til kernen, kræver det derfor, at låget til de sociale mediers algoritmer åbnes, og at forskere får adgang til blandt andet direkte at køre eksperimenter på platformene.

Et første skridt på vejen er et storstilet samarbejde mellem Meta, selskabet bag Facebook og Instagram, og en lang række forskere, der har fået lov til at køre eksperimenter på platformene under den amerikanske præsidentvalgkamp i 2020. Hvad angår Facebook og Instagrams evne til at fremhæve politisk polarisering hos deres brugere under præsidentvalgkampen, er resultaterne tydelige. Således afskar forskerne en gruppe vælgeres adgang til Facebook og Instagram under valgkampen, for en anden gruppe skruede de op og ned på algoritmen i forhold til mængden af indhold fra nyhedssider, som vælgerne ideologisk var på linje med, for en tredje gruppe fjernede de muligheden for at se delinger (*reshares*) og for en fjerde gruppe fjernede de den algoritmiske kuration af indholdet på *newsfeed*. Ingen af eksperimenterne viste tegn på, at ændringerne havde en betydning for brugernes grad af politisk polarisering (Guess et al., 2023a; Guess et al., 2023b; Nyhan et al., 2023; Allcott et al., 2024).

Studierne fra samarbejdet med Meta er ikke uproblematiske, ej heller overkommer de problemet med, at vi ikke kan finde en verden, der ikke er sølet ind i sociale medier, hvorfor studierne alene indikerer, at indhold på Facebook og Instagram ikke synes at polarisere nu, hvor vi alle har sociale medier og har brugt dem i rigtig mange år. Hvorom konklusionen derfor er, at forskningen på dette tidspunkt har svært ved at finde, at sociale medier polariserer brugerne, er der brug for mere forskning for endegyldigt at svare på bekymringen.

Sociale mediers betydning for trivsel

Sociale medier er for tiden et hedt emne både politisk og forskningsmæssigt. Globalt, såvel som i Danmark, er der et stort fokus på mængden af tid, befolkningen, og særligt unge, bruger på sociale medier. Frygten er, at de sociale medier trækker et spor med sig af dårligdomme, deriblandt særligt mistrivsel, som afsnittet her fokuserer på.

Om end spørgsmålet har givet anledning til analyser og diskussioner i forskningsverdenen i årevis (se bl.a. Orben & Przybylski, 2019 og Twenge et al., 2020) nåede spørgsmålet kogepunktet i foråret 2024, hvor den kendte amerikanske socialpsykolog, Jonathan Haidt, udgav sin seneste bog *The Anxious Generation*. Haidt argumenterer for, at frygten for sociale medier er berettiget: mobiltelefonen og sociale medier har erstattet den fysiske leg og socialisering og medført markante mentale sundhedsproblemer, fordi unge piger konstant ser usunde skønhedsidealer og unge drenge bliver eksponeret for voldsomt indhold såsom mord og porno. Konkret viser Haidt, hvordan stigningen i mentale sundhedsproblemer kommer samtidig med at flere bruger digitale teknologier i stigende grad.

Haidts argumenter har fået mange til tasterne (Odgers, 2024; Thorp, 2024) med henvisning til en række forskningsmæssige resultater, der ikke kan finde klar evidens for den kritiske fremstilling af sociale medier, når der kigges på sammenhængen over tid (Heffer et al., 2019; Odgers & Jensen, 2020; Orben, 2020; Valkenburg et al., 2022). Direkte modsat argumentet fra Haidt, så mener nogen at sammenhængen er omvendt, hvor unge, der allerede har mentale sundhedsproblemer, bruger sociale medier mere og måske på andre måder end unge uden mentale sundhedsproblemer (Heffer et al., 2019). På samme måde finder studier, der eksperimentelt forsøger at teste tesen ved at få en gruppe til at afholde sig fra sociale medier i en periode, at det ikke er muligt i gennemsnit at finde en effekt af sociale medier på trivsel, der er forskellig fra nul (Ferguson, 2024; Radtke et al., 2022).

Forskningsresultaterne er dog til stor debat. Der er således flere studier, der peger i modsat retning og derved på sociale mediers negative betydning for særligt børn og unges mentale helbred (Weigle & Shafi, 2024; Blanchard et al., 2023; Khalaf et al., 2023; Ergün et al., 2023). Ligeledes har nogle studier forsøgt sig med forskningsdesigns, der udnytter den naturlige variation i bl.a. den gradvise udrulning af det sociale medie, Facebook. I og med, at Facebook blev udrullet gradvist på amerikanske universiteter, kunne forskere sammenligne de studerendes trivsel før og efter Facebook blev udrullet på deres campus (Braghieri et al., 2022). Her finder forskerne, at Facebook havde en negativ indvirkning på mental sundhed, med argumentet om, at Facebook medfører mulighed for på en usund måde at sammenligne sig med sine medstuderende.

Det omtalte studie møder dog også kritik (Eckles, 2023) og på et højere abstraktionsniveau kan det være svært altid at vide, præcis hvilket ben man bør stå på. For det første kan uenighederne – også de, der ikke er strengt forskningsmæssige – handle om, hvordan trivsel defineres og at vi ved meget lidt om, hvad det egentligt er for noget indhold folk ser på deres skærme eller hvordan sociale medier bliver brugt forskelligt af forskellige personer (Reeves et al., 2020). For det andet er det svært at isolere effekten af sociale medier, når de er så udbredte og integrerede dele af de flestes dagligdag, hvorfor de fleste studier kun er i stand til at vise en delmængde af den samlede fortælling. At forskningen holder snuden i sporet og får fat i mere og bedre data er derfor afgørende for at få et præcist billede af problemets omfang.

Fortsat ubesvarede spørgsmål og forslag til forskningsdesigns

Hvor den forrige del af rapporten fokuserede på, hvad vi fra forskningen ved om misinformation, politisk polarisering og trivsel, skifter vi nu fokus til, hvilke spørgsmål der endnu står ubesvarede hen og hvordan det vil være muligt at svare på nogle af dem i fremtidige rapporter. I det følgende er der derfor først en gennemgang af en række ubesvarede spørgsmål og dernæst en gennemgang af tre forskningsdesigns, der vil kunne svare på nogle af de spørgsmål.

Fortsat ubesvarede spørgsmål

De fortsat ubesvarede spørgsmål i dette afsnit bliver formuleret som en række forskningsspørgsmål, hvoraf nogle tages op i det efterfølgende afsnit for at definere forskningsdesigns, der kan anvendes i fremtidige rapporter.

For forskningsspørgsmålene, der gælder misinformation, formulerer vi desuden en række spørgsmål relevante i lyset af **generativ AI.** Hvor litteraturgennemgangen på nogle stræk nedtoner frygten for misinformation, er der fare for, at generativ AI kan ændre på det. Men der er også potentielt mulighed for positive bivirkninger. Dette sætter vi skarpt på med spørgsmål, der indfanger potentielle skift i spredning og tro på misinformation som følge af generativ AI.

Generativ AI (generativ kunstig intelligens) er teknologier, der kan skabe nyt indhold (fx tekst, billeder, lyd eller video) på baggrund af eksisterende data og instruktioner. Det mest kendte eksempel på generativ kunstig intelligens er chatbotten ChatGPT

Misinformation

Danskernes adfærd med misinformation

Litteraturgennemgangen viste to slående træk. For det første, at hvad vi fra forskningen ved om misinformation i overvejende grad er amerikansk. USA er imidlertid meget forskellig fra Danmark, hvad angår sprog, polarisering og (politisk) kultur – for blot at nævne få vigtige forskelle. Det kan derfor være svært direkte at oversætte resultaterne fra USA til Danmark, hvorfor det er nødvendigt at vide mere om misinformation i en dansk kontekst. For det andet er langt det meste eksisterende forskning baseret på spørgeskemaundersøgelser, der således alene er i stand til at identificere folks holdninger og opfattelser af misinformation og sjældent deres adfærd.

De første forskningsspørgsmål, vi derfor stiller i afsnittet her, betoner derfor behovet for at indfange danskernes adfærd med misinformation.

I hvilket omfang bliver danskerne eksponeret for og interagerer med misinformation?
I hvilket omfang påvirker danskeres interaktion med misinformation deres generelle informationsdiæt på sociale medier?

Produktion og distribution af (mis)information
Én af de helt centrale virkninger af generativ AI er ikke blot at skabe utilsigtet misinformation gennem "hallucinationer", men også muligheden for at producere meget og virkelighedstro information, hurtigt. Derfor er en af farerne, at generativ AI kan bruges til bevidst at skabe misinformation i store mængder.

Dette er særligt vigtigt, givet at opbygningen af sociale medier (i varierende omfang) er bygget op omkring viralitet: et opslag, et billede eller en video, der bryder gennem loftet og bliver spredt til millioner af mennesker. Om end processen ikke er tilfældig, er formlen heller ikke nødvendigvis klar for andre end algoritmen, der styrer showet, selv. Med generativ AI, bliver det lettere for ondsindede aktører at skabe mere indhold, der alt sammen indgår som lodder i det store viralitetslotteri, og dermed øger risikoen for, at misinformation går viralt.

På samme tid kan generativ AI naturligvis også bruges til at både skabe og cirkulere sand information, hvorfor balancen mellem hvor mange lodder gode og ondsindede aktører har i lotteriet, ikke nødvendigvis ændres.

I hvilket omfang gør generativ AI det lettere at producere og distribuere misinformation?

Effekterne af (mis)information
En af de situationer, hvor der er høj efterspørgsel på information, er i krisesituationer. De tilfælde er imidlertid kendetegnet ved, at der ofte ikke er meget information at hente, fordi medier, politikere og andre fortsat er i gang med at undersøge sagen.

Befolkningens viden om generativ AI's eksistens og dens risiko for at blive brugt til at sprede misinformation, risikerer derfor at sænke tilliden til de nyheder vi ser og bliver eksponeret for i krisesituationer.

Omvendt kan befolkningens opmærksomhed herpå også være med til at øge efterspørgslen efter autentisk viden ved at være mere kritisk og reflekteret over de nyheder, de præsenteres for.

Hvordan påvirker generativ AI befolkningens tillid til nyheder og information i krisesituationer?

Ét af de mest omdiskuterede tilfælde af generativ AI i de senere år har været såkaldte *deepfakes*. *Deepfakes* er kendetegnede ved at erstatte en persons udtryk med et andet på en overbevisende måde. Det kan være at skabe en video, hvor en politiker giver udtryk for holdninger og udsagn, vedkommende aldrig har haft eller sagt. Et eksempel på dette så vi i foråret 2024, hvor Dansk Folkeparti kom i vælten for at have produceret og delt en *deepfake* med Mette Frederiksen, der gav udtryk for at danskernes arbejder for lidt og at samtlige helligdage derfor afskaffes (Uldall, 2024).

Udover at være problematisk i sig selv, risikerer sådanne fænomener at føre til, hvad der kaldes *tainted truth*. Situationer, hvor der hos folk sås tvivl om, hvorvidt de kan stole på ellers sand information med frygt for, at de ikke er i stand til at skelne mellem, hvad der er sandt og hvad der er falsk.

Modsat er det muligt, at diskussionen om misinformation og generativ AI øger befolkningens kritiske sans. Lig idéen om, at kunne "vaccinere" mod misinformation, vil en sådan diskussion gøre folk bedre til at identificere, hvornår information er til at stole på, og hvornår den ikke er.

I hvilket omfang påvirker generativ AI befolkningens evner og tro på, at de kan skelne mellem sand og falsk information?

Afsnittet om udbredelsen og troen på misinformation berørte, hvilke grupper af befolkningen som forskningen har haft udpeget til at være de mest udsatte når det gælder misinformation. Et af argumenterne for, hvorfor nogle er mere udsatte for at falde for misinformation handler om, at f.eks. ældre ikke er lige så vante på internettet som den yngre del af befolkningen.

Med generativ AI bliver det lettere at producere højkvalitetsindhold i en skala, hvor det i mindre grad er nødvendigt at fokusere elefantbøssen på enkelte grupper. I stedet kan det blive lettere med spredehagl at gå efter at overbevise alle befolkningsgrupper. På den positive side, kan det også blive lettere at nå ud til folk med troværdig information om hvilken information man skal være påpasselig overfor.

I hvilket omfang udviskes forskelle mellem dem, som er gode og mindre gode til at identificere misinformation som følge af generativ AI?

Interventioner mod misinformation

En vigtig indsats i kampen mod misinformation er interventioner såsom advarselstegn og faktatjek. Sådanne interventioner risikerer imidlertid at blive påvirket af generativ AI, der blandt andet kan oversvømme *crowdsourcede* interventioner med falske faktatjek. Sådan oversvømmelse kan medføre såkaldt *warning fatigue* (Morrow et al., 2022), hvor den store mængde af advarsler gør folk trætte af, at der er advarsler og faktatjek på alting. Ligeledes bliver faktatjek sværere at foretage, af den grund at der er større usikkerhed om, hvorvidt et billede eller video er AI-genereret eller ikke. Dette åbner døren for, at ondsindede aktører kan snyde sig rundt om kritik under henvisning til, at informationen om dem er falsk og AI-genereret.

Selvom rapporten her ikke har fokuseret på den del af forskningen, der beskæftiger sig med identifikation af misinformation, er dette et punkt, hvor generativ AI indebærer positive muligheder. Generativ AI kan potentielt bruges til at skalere ellers svært-skalerbare interventioner og gøre det lettere at identificere misinformation.

I hvilket omfang påvirker generativ AI effekten af interventioner mod misinformation?

Politisk polarisering

Noget af det forskningen fortsat ikke kan fortælle os meget om, er hvordan politisk polarisering på sociale medier ser ud uden for det meget elite- og affektiv polariserede USA. Alene det, at udbrede fokus til at omfatte flerpartisystemer som Danmark, vil derfor øge samfundets viden om det potentielt polariserende liv på de sociale medier. På det bagtæppe finder vi det relevant at fokusere på spørgsmålene:

I hvilket omfang bliver danskerne præsenteret for forskelligt indhold i forhold til hinanden på sociale medier?
I hvilket omfang bliver indholdet mere polariseret over tid?
I hvilket omfang interagerer danskerne med polariserende indhold?

Trivsel

Om end det ikke er let at isolere effekten af sociale medier på brugernes trivsel, er der fortsat vigtige og mere tilgængelige spørgsmål at undersøge, særligt i Danmark. Der mangler fortsat viden om, hvor meget problematisk indhold børn og unge eksponeres for, og hvilke forskelle der er i det indhold unge, der henholdsvis trives og mistrives, ser på sociale medier. Det er derfor oplagt at teste de hypoteser, der ofte fremsættes i dokumentarer, politiske debatter og lignende, navnligt ved at undersøge spørgsmålene:

I hvor høj grad eksponeres unge for problematisk indhold på sociale medier?
Hvad kendetegner forskelle i interaktion- og eksponeringsmønstre mellem unge, der varierer i selvrapporteret trivsel?

Spørgsmålene formuleret i dette afsnit har peget på nogle af de områder, hvor vi fortsat mangler viden indenfor emnerne misinformation, trivsel og politisk polarisering i relation til sociale medier. Det er derfor oplagte spørgsmål for fremtidig forskning og undersøgelser at tage op. I næste afsnit udplukker vi nogle af dem og udfolder, hvordan de vil kunne undersøges.

Forslag til fremtidige temaer og forskningsdesigns

Den primære udfordring når det gælder undersøgelser af sociale medier og deres betydning for samfundet er manglende data og gode forskningsdesigns. Dette afsnit opstiller tre forskningsdesigns på baggrund af nogle af de uafdækkede forskningsspørgsmål udlagt i forrige afsnit. De tre forskningsdesigns udfolder, hvordan det er muligt at undersøge de forskellige spørgsmål og hvad det konkrete databehov til de enkelte undersøgelser er. Disse forskningsdesigns kan bruges som pejlemærker for fremtidige undersøgelser.

For at holde fokus på Medieaftalens prioriteter for den uvildige undersøgelse, opstiller vi tre forskningsdesigns, der fokuserer på henholdsvis danskernes adfærd med misinformation online, politisk polarisering på sociale medier i Danmark, og sociale mediers betydning for danske børn og unges trivsel.

Forskningsdesign I: Danskernes adfærd med misinformation online

Litteraturgennemgangen fokuserede primært på folks opfattelser af misinformation. Årsagen hertil er, at folks adfærd ift. misinformation er data, der ikke er let tilgængelig. Det er dog et vigtigt spørgsmål, og dette forskningsdesign optegner derfor, hvordan en undersøgelse af danskernes adfærd med misinformation på sociale medier kunne se ud. Forskningsspørgsmålene er som følger:

- I hvilket omfang bliver danskerne eksponeret for og interagerer med misinformation?

- I hvilket omfang påvirker danskeres interaktion med misinformation deres generelle informationsdiæt på sociale medier?

Forskningsdesignet kræver data om, hvilket indhold danskerne eksponeres for og interagerer med på forskellige platforme, samt viden om deres medie- og informationsdiæt, ideologiske holdninger, prædispositioner og demografiske træk. Da forskningsdesignet har alle danskere for øje er Facebook, Instagram, og YouTube de ideelle platforme at fokusere på, da det er disse platforme, der er de mest udbredte i Danmark og har været det i lang tid. Specifikt vil data til dette design komme fra to kilder:

1) **Spørgeskemaundersøgelse af et repræsentativt udsnit af danskere**, indeholdende spørgsmål til ideologisk selvplacering, tro på konspirationsteorier, prædispositioner, informationsdiæt og en række demografiske variable.
2) **Observationel adfærdsdata om indholdseksponering gennem DSA.** Data skal indeholde hvilket indhold det repræsentative udsnit af danskere har været eksponeret for og interageret med på Facebook, Instagram, YouTube og Snapchat. Derudover skal data indeholde information om, hvorvidt indholdet har været indrapporteret som misinformation og hvorvidt opslaget senere er blevet taget ned af platformen.

Dataen kan bruges til en række analyser. For det første **deskriptive opgørelser over, hvor meget og hvilken type misinformation, danskerne eksponeres for og interagerer med**. Dette gælder både mængden af misinformation og andelen det udgør af den samlede informationsdiæt, ligesom det kan danne overblik over, hvilke kilder misinformationen oftest stammer fra. For det andet **hvorvidt danskerne i højere grad eksponeres for misinformation efter de har interageret med det**. Dette kan eksempelvis undersøges med et *staggered difference-in-differences* design, hvor det undersøges, hvorvidt interaktion med misinformation forårsager stigning i fremtidig eksponering for misinformation. For det tredje **korrelationer mellem danskernes baggrundskarakteristika og eksponering for misinformation**. Dette kunne være, hvorvidt nogle danskere i højere grad bliver eksponeret for misinformation, f.eks. på tværs af ideologisk selvplacering, tro på konspirationsteorier eller politisk tillid.

Forskningsdesign II: Politisk polarisering på sociale medier i Danmark

Langt størstedelen af den eksisterende forskning på spørgsmålet om politisk polarisering er foregået i USA. USA er imidlertid som nævnt meget anderledes fra Danmark på en række centrale områder, deriblandt niveauet af elite polarisering. Det er derfor svært direkte at overføre sådan viden til Danmark. Konkret er forskningsdesignet her derfor opstillet med henblik på at kunne svare på følgende forskningsspørgsmål:

- I hvilket omfang bliver danskerne præsenteret for forskelligt indhold i forhold til hinanden på sociale medier?
- I hvilket omfang bliver indholdet mere polariseret over tid?
- I hvilket omfang interagerer danskerne med polariserende indhold?

For at kunne besvare de to forskningsspørgsmål er der behov for observationel data af, hvilket indhold danskere ser og interagerer med på sociale medier, samt hvordan dette indhold udvikler sig over tid. Til dette forskningsdesign er Facebook og Instagram de mest oplagte

sociale medier at undersøge data fra, da det er de to sociale medier, den største andel af danskere bruger. Specifikt vil data til dette design komme fra tre kilder:

1) **Spørgeskemaundersøgelse blandt et repræsentativt udsnit af danskere**, indeholdende spørgsmål til ideologisk selvplacering, affektiv polarisering og en række demografiske variable.
2) **Datadonationer fra samme repræsentative udsnit af danskere**, indeholdende hvilke sider de har interageret med på Facebook og Instagram, samt hvilke interesser platformene har udledt på baggrund af deres adfærd.
3) **Observationel data om indholdsskabelse gennem *Meta Content Library***. Her er det muligt at hente alle opslag fra offentlige sider (f.eks. medier og politikere) på Facebook siden 2009.

Denne data kan bruges til en række analyser. For det første kan med ved brug af **netværksanalyse undersøge danskernes afstand til hinanden** ift. hvilke sider de følger og interagerer med. Dette vil være med til at beskrive, hvor stor afstand, der er mellem forskellige befolkningsgrupper ift., hvad de ser og interagerer med på sociale medier. For det andet **korrelationer mellem, hvad danskerne interagerer med og politiske overbevisninger,** hvilket kan identificere, hvor stor grad af politisk og indholdsmæssig segregering, der er på sociale medier. For det tredje **tidsserieanalyse af graden af polarisering i indhold fra 2009 til i dag**. Dette vil kunne undersøges ved at analysere politikere, meningsdannere og interesseorganisationers delinger af nyheder på sociale medier (Eady et al., under udgivelse), ligesom det vil være muligt at undersøge udviklingen i negativ indhold over tid med sentimentanalyse og analyse af danskernes type af reaktioner på indhold.

Forskningsdesign III: Social mediers betydning for danske børn og unges trivsel

Som litteraturgennemgangen beskrev, er det fortsat omdiskuteret i hvor høj grad unge eksponeres for problematisk indhold på sociale medier og i hvor høj grad det påvirker dem (positivt og negativt). Da vi fortsat mangler klare teoretiske forventninger om blandt andet forskelle mellem unge i deres eksponering for problematisk indhold og dets effekter, vil forskningsdesignet her fokusere på følgende to mere grundlæggende forskningsspørgsmål:

- I hvor høj grad eksponeres unge for problematisk indhold på sociale medier?
- Hvad kendetegner forskelle i interaktion- og eksponeringsmønstre mellem unge, der varierer i selvrapporteret trivsel?

For at kunne besvare de to forskningsspørgsmål er der behov for observationel data om, hvilket indhold unge bliver eksponeret for på sociale medier samt selvrapporterede data om unges trivsel. For at få det stærkest mulige grundlag undersøges data fra Instagram, TikTok og Snapchat, der er særligt udbredte blandt unge i Danmark. Data vil i dette tilfælde komme fra to kilder:

1) **Spørgeskemaundersøgelse blandt et repræsentativt udsnit unge**, deriblandt indeholdende spørgsmål, der måler selvrapporteret trivsel, samt sociale og demografiske faktorer såsom interaktion med venner, køn, alder og forældres uddannelsesstatus.
2) **Observationel adfærdsdata om indholdseksponering gennem DSA**. For hver af de repræsentativ udvalgte unge bedes platformene give adgang til data for, hvilket indhold de unge har været eksponeret for og interageret med de seneste to år samt

information om, hvilke interesser platformene har infereret den unge til at have på baggrund af deres adfærd.

Denne data kan danne grundlag for en række analyser. For det første **deskriptiv præsentation af problematisk indhold** baseret på indholdskodning, der tager højde for den totale mængde af indhold præsenteret. For det andet **korrelation mellem trivsel og eksponering for problematisk indhold**. For det tredje **netværksanalyse af unges interaktionsmønster**, der giver indblik i hvilke interesser og interaktioner, der korrelerer med problematisk indhold.

Begrænsninger og ressourcekrav for de tre forskningsdesigns

De beskrevne forskningsdesigns er opstillet til at kunne give en mere systematisk deskriptiv analyse end hvad forskningslitteraturen indtil videre generelt har haft mulighed for. Det er imidlertid relevant at notere, at forskningsdesignsene ikke er i stand til at afgøre kausale forhold. F.eks. kan forskningsdesignet om forholdet mellem problematisk indhold på sociale medier og unges trivsel ikke adskille, hvad der skyldes, at unge med varierende trivsel opsøger forskelligt indhold, og hvad der skyldes, at unge mistrives, fordi de eksponeres for problematisk indhold. Et sådan design er svært at opstille uden eksperimentel manipulation, hvilket hverken er muligt med eksisterende datakilder og i de fleste tilfælde ikke vil være etisk forsvarligt. De tre forskningsdesigns vil derfor primært være udtryk for deskriptiv inferens kombineret med enkelte kvasi-eksperimentelle designs.

Alle tre forskningsdesign er ressource- og tidskrævende. Data, der enten indhentes gennem mulighederne i forordningen om digitale tjenester (DSA) eller *Meta Content Library* kræver længerevarende og usikre processer med dataanmodninger. Derudover kræver datadonationer en relativ stor indsats for deltagerne, da de skal ind og rekvirere data fra den pågældende platform, vente på at modtage data og dernæst *uploade* data til forskningsgruppen, hvilket alt andet lige øger prisen for at indhente data. Ligeledes er spørgeskemaundersøgelser blandt børn og unge dyrere, da vi har med en gruppe at gøre, der er sværere at få fat i til den type undersøgelser. Alle tre forskningsdesigns kræver derfor både tid og betydelige ressourcer, hvis de skal udføres i det beskrevne format.

Ny undersøgelse af danskernes syn på og evner ift. misinformation og generativ AI

I denne sidste del af rapporten præsenterer vi ny viden om danskernes syn og evner ift. misinformation og generativ AI. Det gør vi på baggrund af en spørgeskemaundersøgelse foretaget i juni 2024, hvor vi har spurgt 2.091 danskere, som udgør et repræsentativt udsnit af befolkningen, om tre forskellige temaer.

Som vi så i rapportens første del, så er indholdet og indholdsmoderation på sociale medier vigtig, når vi taler om den offentlige samtale online, og det er et område, hvor der fortsat er stor kritik af, at der bliver gjort for lidt. Derudover er det et område, hvor kulturelle skismaer bliver tydelige: danskerne, der deltager og diskuterer på sociale medier, er i de fleste tilfælde underlagt amerikanske kulturelle normer, der til tider kolliderer med det generelle danske frisind. Det første tema, vi har bedt danskerne forholde sig til er derfor deres *holdninger til indholdsmoderation og regulering af sociale medier*.

Litteraturgennemgangen viste, hvordan det er usikkert, hvor gode folk er til at skelne mellem sand og falsk information, og hvordan denne forskel potentielt kan blive større i lyset af generativ AI. Der er imidlertid ingen af sådanne undersøgelser i Danmark, ligesom de fleste eksisterende undersøgelser lider under den mangel, at de ikke er i stand til at adskille folks hukommelse og evne til at identificere misinformation. Det andet tema, vi har bedt danskerne forholde sig til, er derfor deres *evner til at identificere misinformation*.

Litteraturgennemgangen viste derudover negative indirekte effekter af misinformation, såsom hvorvidt frygten for misinformation også sænker tiltroen til sande nyheder. Ligeledes var det en konklusion, at andelen af misinformation online i forhold til anden information er mindre end de fleste nok går og tror. Derfor er vi interesseret i at vide, hvorvidt opmærksomheden og omtalen af misinformation potentielt har negative afledte effekter, der er vigtige at tage højde for. Det tredje og sidste tema, vi derfor har bedt danskerne forholde sig til er *hvordan de påvirkes af forskellige udlægninger af faren ved misinformation*.

Holdninger til indholdsmoderation og regulering på sociale medier

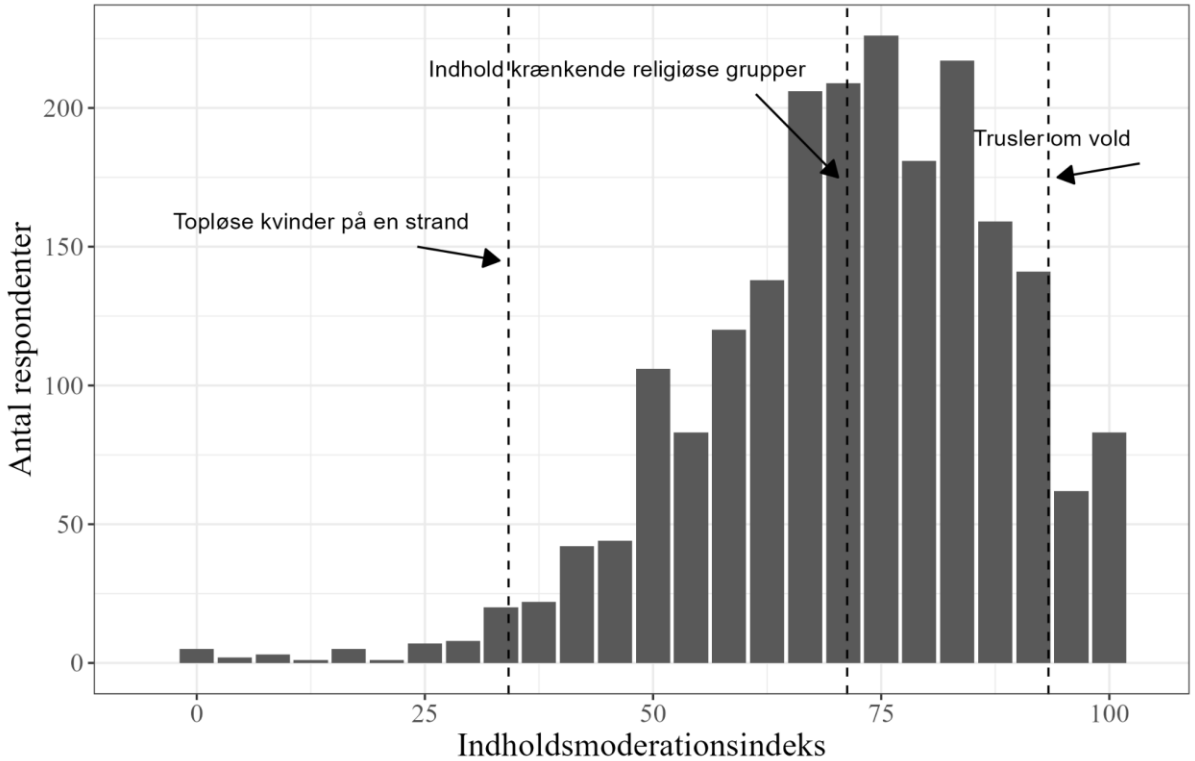
Det første tema vi behandler her, er spørgsmålet om danskernes holdning til indholdsmoderation og regulering af sociale medier. Litteraturgennemgangen viste imidlertid, at det ikke nødvendigvis er holdninger, der er lette at indfange. Folk har nemlig en tendens til at have modstridende ønsker, som på samme tid at ønske stor grad af indholdsmoderation og stor grad af ytringsfrihed (Morrow et al., 2022).

Vores spørgeskemaundersøgelse tager i stedet fat i danskernes holdning til faktiske indholdsmoderationspolitikker på Facebook. Vi udplukker forskellige eksempler på indhold, der ifølge platformens egen politik bliver fjernet fra platformen, og spørger respondenterne, i hvor høj grad de er enige i, at den type indhold bør fjernes fra sociale medier. Disse eksempler varierer også i ekstremitet, hvor det mest ekstreme spørgsmål, holdningen til om trusler om vold, er decideret ulovligt ifølge dansk ret (Straffeloven § 266). Tilsammen giver svarene os indikationer på danskernes overordnede tilfredshed med eksisterende indholdsmoderationspolitikker og forskelle på tværs af forskellige typer af politikker. Alle

præsenterede eksempler og korresponderende udsagn fra Facebook's indholdsmoderationspolitik findes i bilag 3A. Det bør dog bemærkes at vi ikke undersøger, hvorvidt Facebook rent faktisk følger sin egen indholdsmoderationspolitik og fjerner indhold i overensstemmelse hermed.

Figur 4 viser fordelingen af respondenternes enighed i præsenterede indholdsmoderations-eksempler. Da hver respondent er blevet bedt om at vurdere seks typer af indhold, er hver respondents holdning til indholdsmoderation samlet i et indeks, hvor en lav værdi angiver, at respondenten ikke er enig i indholdsmoderationspolitikkerne, mens en høj værdi angiver stor enighed. Derudover er de to typer af indhold, som henholdsvis flest og færrest er enige i bør fjernes, markeret med stiplede linjer ved deres respektive gennemsnit, ligesom der også er en markering af indholdstypen, der ligget tættest på fordelingens gennemsnit.

Figur 4: Danskerne er generelt – men ikke helt – tilfredse med eksisterende indholdsmoderationspolitikker



Præcise spørgsmålsformuleringer:
- Trusler om vold, der kan føre til alvorlig skade
- Billeder af topløse badende kvinder på en strand, som er lagt op med samtykke fra kvinderne
- Indhold, der omtaler religiøse grupper som eksempelvis dumme eller idioter

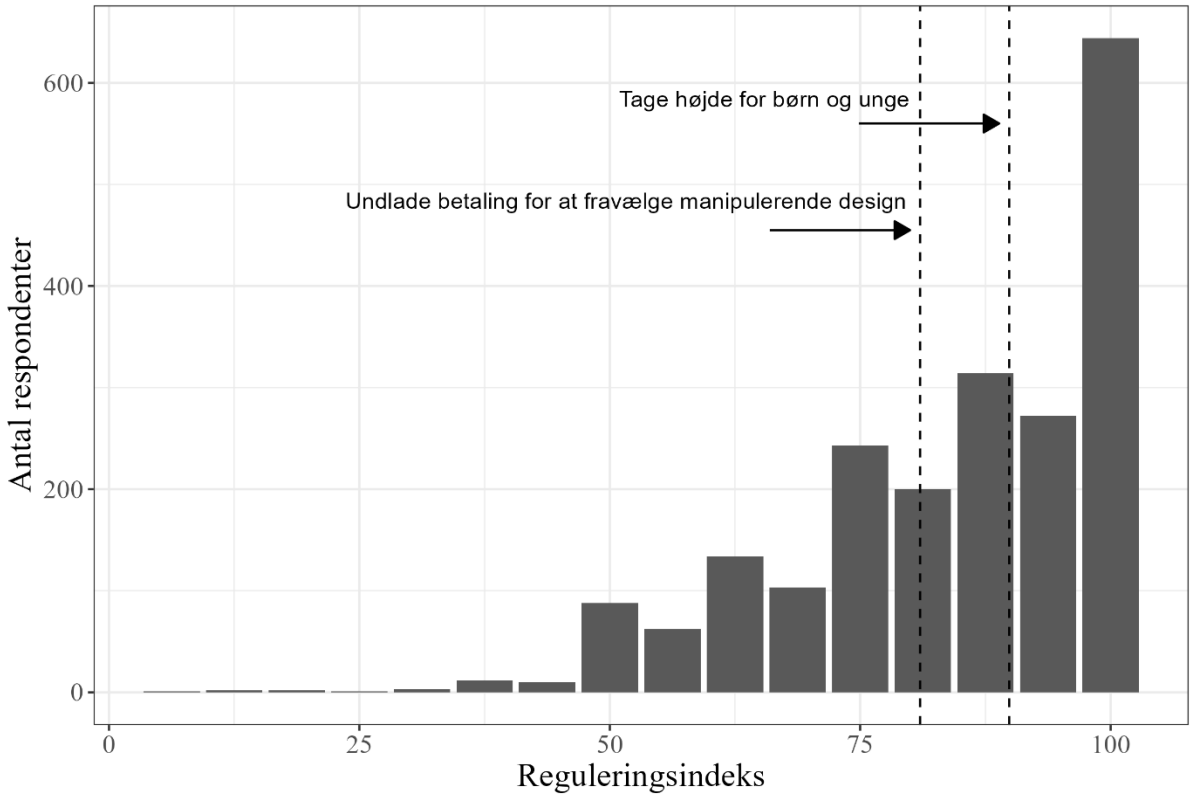
Figur 4 viser, hvordan **danskerne generelt er enige i den indholdsmoderation de oplever på Facebook, men ikke ubetinget**. For det første viser spredningen i fordelingen, at der er stor forskel på, hvor enige danskerne er. Men det er også værd at bemærke, at de fleste danskere har moderat positive holdninger til eksisterende indholdsmoderation. Blot fire procent af danskerne erklærer sig "meget enig" i alle politikkerne og kun otte procent er i gennemsnit mere uenige end de er enige. For det andet er der stor forskel på, hvor enige danskerne er i de forskellige typer af indholdsmoderation. Næsten alle danskere er meget

enige i, at trusler om vold bør fjernes, mens de fleste er uenige i, at billeder, der viser topløse kvinder på en strand lagt op med samtykke fra kvinderne, bør fjernes. Dette gælder både mænd og kvinder, om end kvinder i gennemsnit på alle spørgsmål er lidt mere positive over for indholdsmoderation end mændene. De politikker, der ligger tæt på fordelingens gennemsnit, er indholdsmoderation af indhold, som krænker religiøse grupper, og oplysninger, som er i modstrid med sundhedsfaglige myndigheders anbefalinger.

Udover indholdsmoderation har vi ligeledes spurgt respondenterne om deres holdninger til fire reguleringsanbefalinger fra regeringens nedsatte tech-ekspertgruppe (Erhvervsministeriet, 2024). Ekspertgruppen kom med en række anbefalinger til, hvordan der sikres demokratisk kontrol med tech-giganternes forretningsmodeller. Disse anbefalinger, ville – hvis de blev gennemført - medføre ændringer i det indhold, vi ser på sociale medier, og hvordan de algoritmisk er opbygget.

Figur 5 viser med samme opbygning som figur 4, fordelingen af respondenternes enighed i de forskellige anbefalinger, hvor en høj værdi på indekset angiver stor enighed, mens en lav værdi angiver lille enighed. De to anbefalinger med henholdsvis flest enige og flest uenige er markeret med stiplede linjer.

Figur 5: Danskerne er generelt meget positivt stemt overfor reguleringsforslag



Præcise spørgsmålsformuleringer:
- tage højde for børn og unge som særligt sårbare grupper, når det gælder manipulation og afhængighed
- undlade at kræve betaling for at brugerne kan fravælge manipulerende design

Figur 5 viser at **danskerne generelt er meget positiv stemt over for anbefalingerne til regulering af tech-giganter**. Knap 1/3 af alle respondenter erklærer sig således "meget enig" i alle fire anbefalinger. Modsat ved vurderingen af indholdsmoderation er der kun lille forskel

i, hvor enig respondenterne er i de enkelte typer af regulering. At tech-giganter skal tage højde for børn og unge som særligt sårbare grupper når det gælder manipulation og afhængighed er flest enige i. Modsat er spørgsmålet om, hvorvidt det bør koste penge at fravælge manipulerende design det med mindst opbakning. Begge spørgsmåls gennemsnit ligger dog højt på indekset, hvilket indikerer to ting. For det første at danskerne generelt synes positivt stemte over for regulering. For det andet at skellene ikke angår hvilken type af regulering danskerne ønsker, men snarere hvorvidt man ønsker mere regulering eller ej.

Danskernes evner til at identificere AI-genereret indhold

Diskussionen om folks evne til at identificere AI-genereret indhold har taget fart i kølvandet på introduktionen af ChatGPT, hvor den brede offentlighed har fået både erfaring med og stor eksponering for generativ AI. En af de demokratiske trusler, der er blevet diskuteret er, hvorvidt billeder, lydklip og videoer skabt med generativ AI kan være med til at sætte ild til spredningen af misinformation og ultimativ snyde folk med virkelighedsnært, men falsk, indhold.

Hvor forskningen, som beskrevet tidligere, har været interesseret i spørgsmålet om, hvorvidt folk er i stand til at identificere misinformation, er der to markante huller i vores forståelse af, hvorvidt folk er i stand til at skelne mellem sand og falsk information. For det første er det tvivlsomt, hvorvidt mange af forskningsresultaterne, der viser folks evne til at *identificere* misinformation eller i stedet folks evne til at *huske*, hvorvidt noget har været skrevet om i medierne. For det andet er der kun lidt forskning i, hvorvidt folk kan skelne mellem ægte billeder og AI genererede billeder.

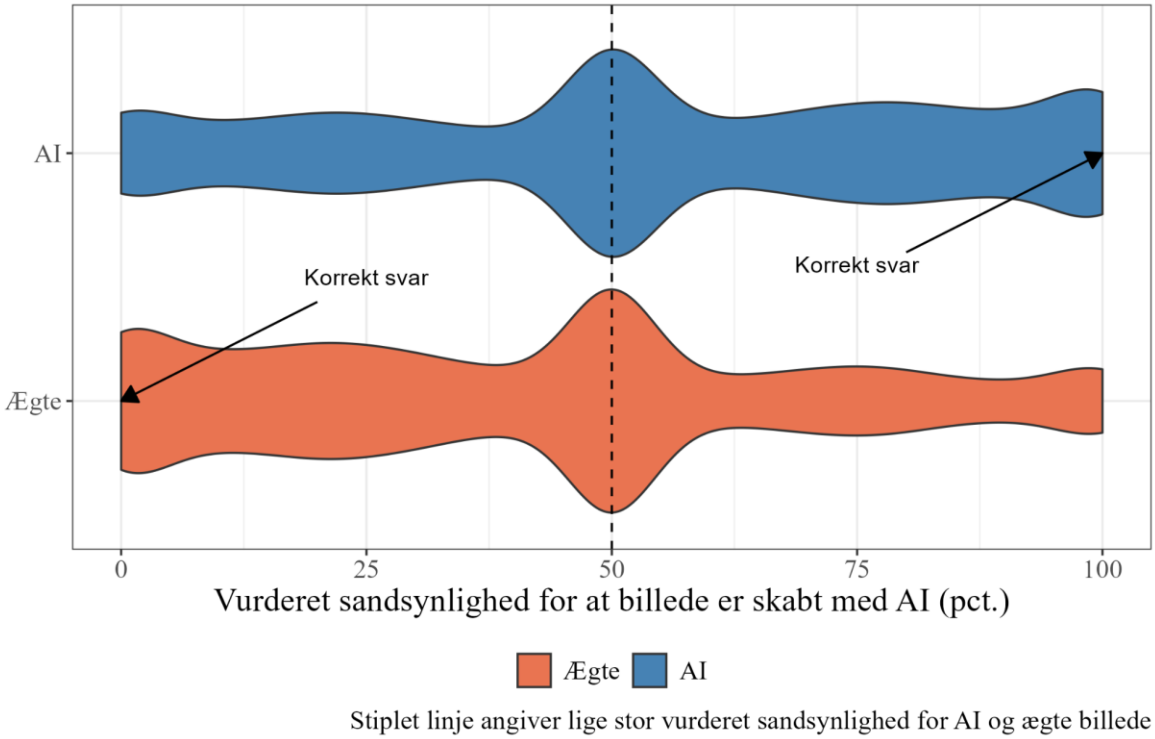
Vi har derfor designet et eksperiment til at kunne vurdere, hvor gode danskerne er til at vurdere ægtheden af de billeder, de bliver eksponeret over for. Hver respondent er blevet bedt om at vurdere syv billeder i forhold til, hvor sikkert de troede, at hver af billederne var skabt med generativ AI på en skala fra 0 til 100. Alle billederne forestillede faktiske begivenheder, hvilket respondenterne ligeledes blev gjort opmærksom på. Det er dog tilfældigt, hvorvidt en respondent blev eksponeret for den faktiske eller AI-genererede version af billedet. En detaljeret beskrivelse af designet kan findes i bilag 3C.

Da generativ AI oftest vil blive anvendt til at fremstille falske scenarier, er det en begrænsning for vores undersøgelse at de scenarier, vi præsenterer respondenterne for, er faktiske begivenheder. Det er dog en nødvendig forudsætning for at kunne have ægte billeder at sammenligne de AI-genererede billeder med, og som derved udelukker effekten af, at en respondent kan huske begivenheden.

Hvordan respondenterne har vurderet ægtheden af de billeder, de har været præsenteret for, ses i figur 6. På figurens x-akse er respondenternes vurdering af, hvor sandsynligt de tror, det er, at billedet er skabt med AI, hvor en værdi på 100 angiver, at respondenten føler sig helt sikker på, at billedet er skabt med AI. Figurens y-akse adskiller de to typer af billeder, der har været præsenteret: henholdsvis ægte og AI-genererede billeder. Jo tykkere fordelingen er, jo flere respondenter ligger der på skalaen.

For det første viser figur 6, at **de fleste danskere ofte føler sig i tvivl om, hvorvidt et billede er skabt med AI**. Det ses ved at fordelingen er tykkest omkring værdien 50, hvilket angiver at respondenten vurderer det lige så sandsynligt, at billedet er skabt med AI som at det er ægte.

Figur 6: Danskerne er ofte i tvivl om billedernes ægthed, men gætter for det meste i retning af det korrekte svar



Stiplet linje angiver lige stor vurderet sandsynlighed for AI og ægte billede

For det andet viser figuren, at **de fleste gæt er i retning af det korrekte svar, men at der er mange gæt, som rammer ved siden af**. Den blå fordeling er tykkere mod højre, som korrekt angiver at billedet er skabt med AI, mens den røde fordeling er tykkere mod venstre, som korrekt angiver at billedet er ægte. Der er imidlertid mange gæt i modsat retning, hvilket indikerer, at mange danskere ikke alene ofte føler sig i tvivl om ægtheden af billedet, men også ofte bliver snydt.

For det tredje fremgår det, at **danskerne synes at være en anelse bedre til at identificere ægte billeder korrekt end AI-genererede billeder.** Det ses ved, at tykkelsen af den blå fordeling i venstre side er tykkere end den røde fordeling i højre side, der hver især angiver det forkerte svar til det præsenterede billede. Det indikerer – om end ikke afgør – at danskerne i mindre grad gætter forkert, fordi de er bange for, at hvad de ser er falsk (*tainted truth*), og i højere grad gætter forkert, fordi AI-skabte billeder ser ægte ud.

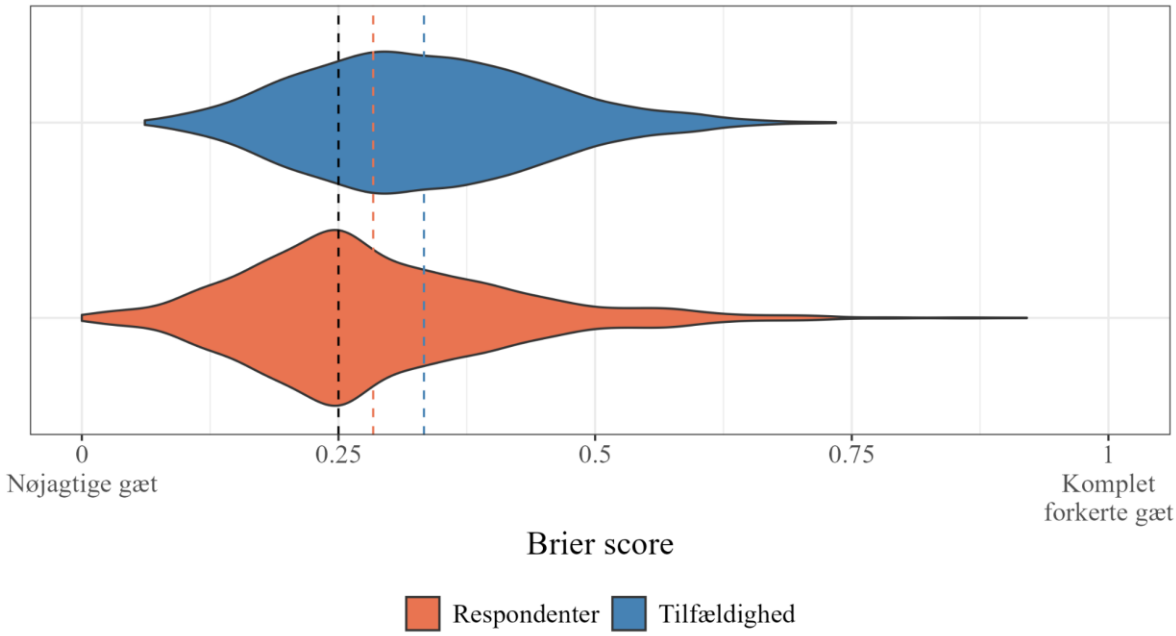
For at kunne evaluere, om danskerne generelt er i stand til at vurdere ægtheden af billeder, anvender vi såkaldte *brier scores*. Fordelen ved denne udregning er, at den bedre vurderer den enkelte respondents samlede evne til at identificere ægtheden af billeder, og er et udbredt mål til at evaluere forudsigelser og lignende. Brier scores udregner forholdet mellem den sandsynlighed respondenten har givet billedet for at være sandt eller falsk, og hvorvidt billedet faktisk er skabt med AI eller ikke. For at forstå logikken i scorerne kan man forestille sig, at det regner 50 pct. af alle dage. Hvis DMI hver aften angiver i sin prognose, at det vil regne

med 50 pct. sandsynlighed næste dag, vil prognosen i gennemsnit være helt korrekt – men ikke særlig nyttig. Det vil være langt mere informativt, hvis DMI hver aften korrekt forudsiger med sikkerhed, om det vil regne næste dag. Brier scores tager højde for både om gæt er korrekte i gennemsnit og om de er udtrykt med sikkerhed, og indfanger derfor denne vigtige sondring.

Hvis DMI i dette tænkte eksempel giver det 70% sandsynlighed for, at det vil regne i morgen, og det så faktisk regner, har DMI været god i sin forudsigelse og får en brier score på 0.09. Regner det ikke, har DMI været knap så god og får en score på 0.49. En brier score på 0 viser således, at en respondent har været fuldstændig sikker i sine svar og har ramt korrekt hver gang, mens en brier score på 1 viser en respondent, der har været fuldstændig sikker i sine svar, men ramt forkert hver gang. I fortolkningen af brier scores er det derfor vigtigt at huske på at *lavere* brier scores er udtryk for *mere nøjagtige* gæt.

Hvor gode respondenterne har været til at identificere AI billeder, kan ses i figur 7. Figuren læses således, at jo tykkere fordelingen er, jo flere respondenter er placeret der på brier scoren og jo tættere på 0 fordelingen er, jo bedre er respondenterne til at identificere billedernes ægthed.

Figur 7: Danskerne klarer sig ikke væsentligt bedre end tilfældighed i vurderingen af billederne



Rød stiplet linje angiver respondenternes gennemsnit.
Blå stiplet linje angiver gennemsnittet ved tilfældige svar.
Sort stiplet linje angiver værdien ved konsekvent 50% vurdering.

For det første viser figur 7, at **danskerne ikke er væsentligt bedre end tilfældige gæt når de skal identificere billedernes ægthed**. Den røde fordeling viser, hvor gode respondenterne i spørgeskemaet var, mens den blå fordeling viser, hvordan resultatet havde set ud, hvis man havde gættet fuldstændig tilfældigt på alle billeder. Som det ses er den røde fordeling en anelse længere til venstre end den blå fordeling, hvilket viser at respondenterne klarer sig lidt bedre, end hvis de bare havde gættet tilfældigt. Dette kan også ses ved de

stiplede linjer, hvor den røde stiplede linje angiver respondenternes gennemsnit, og den blå tilfældighedens gennemsnit. Dog ligger gennemsnittet for respondenterne en anelse over værdien 0,25 (markeret med sort stiplet linje), hvilket er den værdi en respondent ville få, hvis man konsekvent vurderede sandsynligheden til 50 procent. Respondenterne havde derfor i gennemsnit klaret sig bedre, hvis de havde undladt at gætte på, hvad de faktisk troede og i stedet valgt den konservative midterposition.

Det er her vigtigt at huske på, at om noget vil vi forvente, at respondenterne i spørgeskemaet klarer sig *bedre* end de ville have gjort i et virkeligt scenarie, når man eksempelvis færdes på sociale medier. Respondenterne er som led i undersøgelsen blevet gjort opmærksomme på, at nogle af billederne er skabt med AI, og de bruger længere tid på at kigge på billederne (median = 7 sekunder) end man umiddelbart ville, havde man set dem i sit Instagram *feed*. Desuden er billederne i undersøgelsen skabt med offentligt tilgængelige billedegenereringprogrammer på relativ kort tid. Vi ville forvente at kampagner med desinformation lavet af professionelle eller statslige aktører ville være af endnu højere kvalitet.

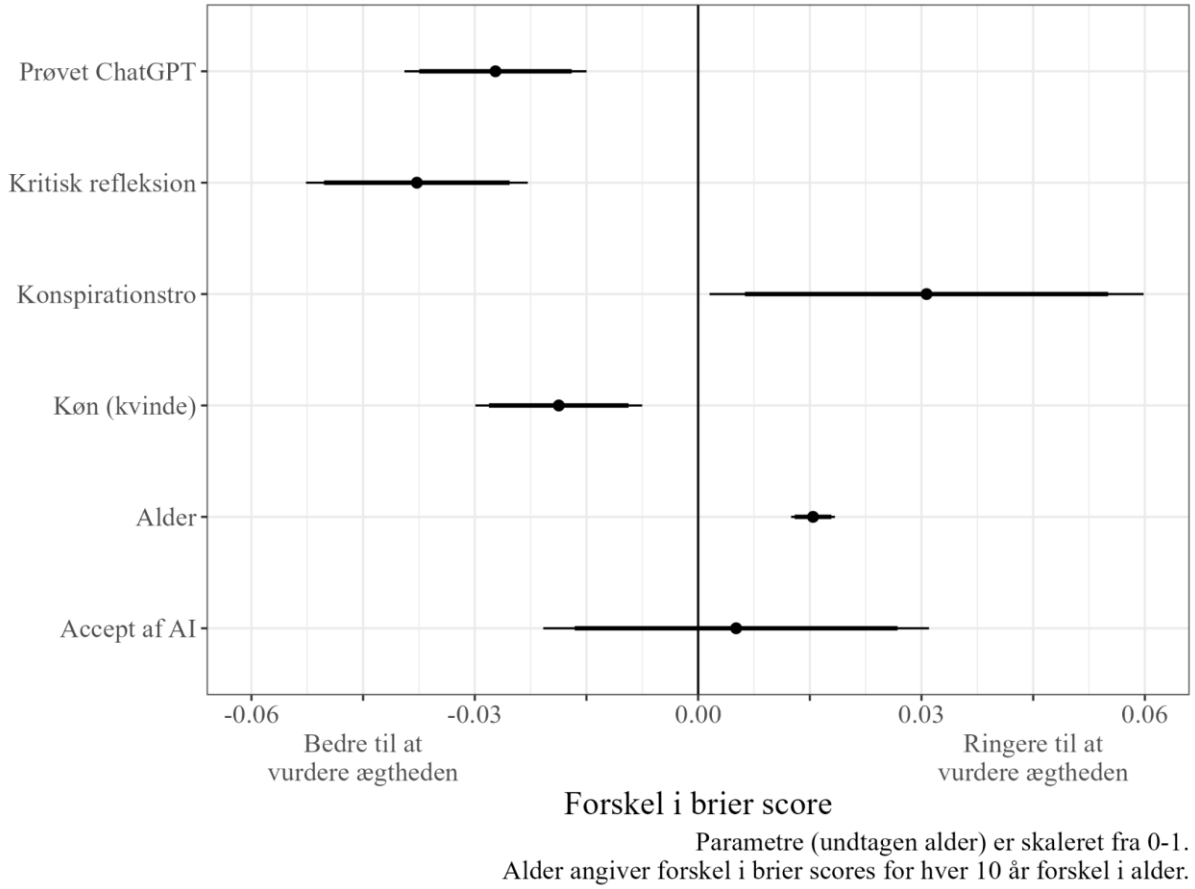
For det andet viser figur 7, at **der er stor forskel på, hvor gode danskerne er til at identificere billedernes ægthed**. Det ses ved, at den røde fordeling strækker sig langt. Der er en mindre gruppe, der ligger tæt på værdien 0, og som derfor er rigtig gode til at identificere billedernes ægthed. Dernæst er der en stor gruppe omkring niveauet med vurderet sandsynlighed på 50 procent, markeret med den sort stiplede linje. Til sidst er der en mellemstor gruppe, der klarer sig ringere, end hvis de havde svaret tilfældigt, hvilket gælder dem, der er at finde til højre for den blå stiplede linje.

Som figur 7 viste, er der stor forskel på, hvor gode respondenterne er til at identificere ægtheden af et billede. Figur 8 viser, hvilke forskelle, der kendetegner de som er gode, og de som er mindre gode til at vurdere ægtheden af billeder. Figuren læses på den måde, at prikken angiver forskellen i brier scores, mens stregerne angiver henholdsvis 90 procent og 95 procents konfidensinterval. Overlapper stregerne 0, markeret med en sort linje, er der ikke med en rimelig sikkerhed forskel på, hvor gode folk er.

Figur 8 viser, hvordan **kvinder, de der har prøvet ChatGPT, og de der har en højere grad af kritisk refleksion, er bedre til at vurdere billedernes ægthed**[7]**.** Samtidig viser figuren hvordan **ældre, og dem som i højere grad tror på konspirationsteorier, er ringere til at vurdere billedernes ægthed**. Der er imidlertid ikke forskel på, hvor gode danskerne er til at afgøre billedernes ægthed afhængigt af, om de synes det er acceptabelt at bruge generativ AI på sociale medier.

[7] Koefficienterne i figur 8 afspejler forskellige modelspecifikationer, som er beskrevet i bilag 3D ud fra et tilhørende *Directed Acyclic Graph* (DAG). Alle koefficienter har samme fortegn og signifikans i bivariate specifikationer, med undtagelse af at "Accept af AI" bliver signifikant og negativ i en bivariat model.

Figur 8: Forskelle i, hvem der er gode til at vurdere ægtheden af billeder



Overordnet indikerer figur 8, at det er bestemte samfundsgrupper, vi skal være nervøse for, når det gælder risikoen for at blive snydt af et AI-skabt billede. Nogle af faktorerne er dog måske at anvende proaktivt: Hvis en person selv har prøvet kræfter med generativ AI og på den måde har en fornemmelse for, hvordan et sådan indhold ser ud, så er det måske muligt at hjælpe resten af befolkningen med at kunne identificere lignende indhold. Der kan blandt andet trækkes en parallel til litteraturgennemgangens fokus på *inoculation*-teknikker, hvor eksponering af f.eks. *deepfakes* i små doser og kontrollerede rammer, måske kan lære folk at identificere de teknikker ondsindede aktører forsøger at bruge til at snyde dem med.

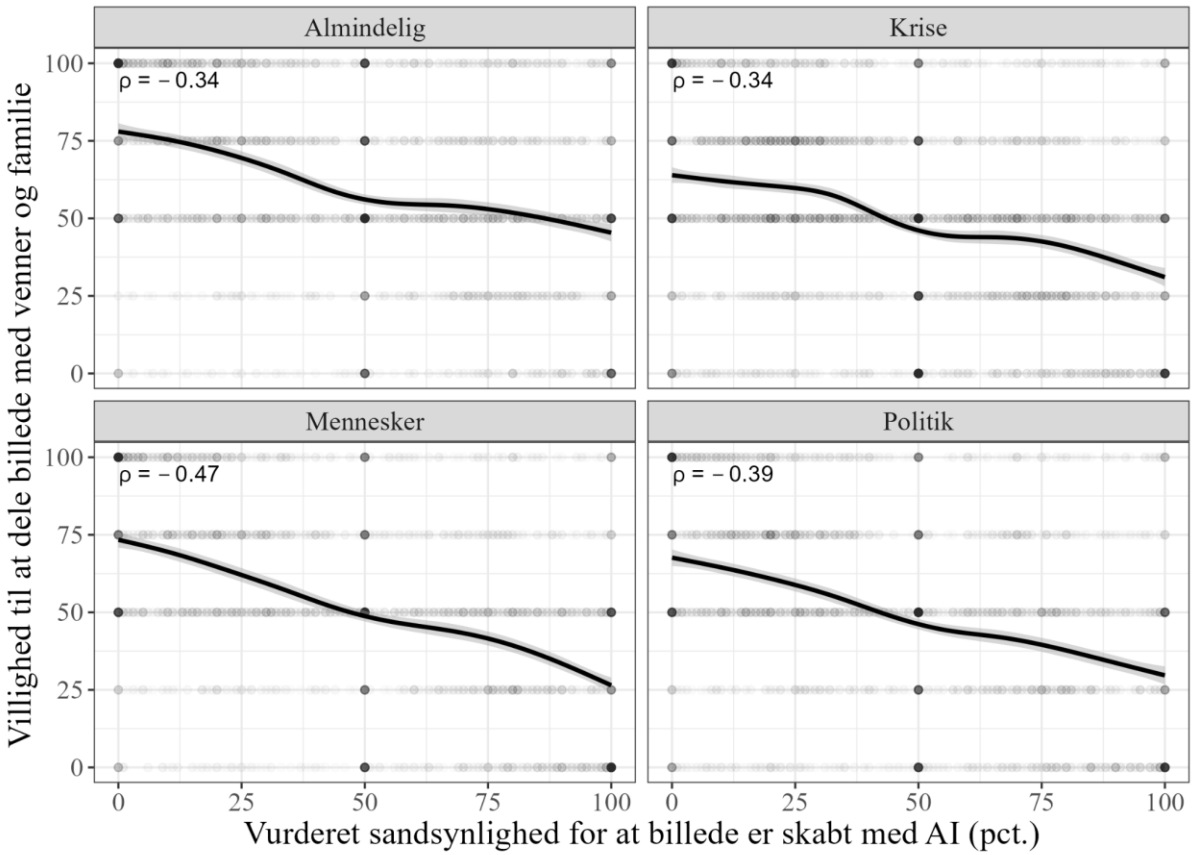
Figuren giver dog også indikationer på, at uddannelse i hvad AI er (ofte under betegnelsen *digital literacy*), og hvordan det kan se ud, ikke er nogen trylledrik. Ældre borgere har måske i kraft af generelt mindre livslæring med IT sværere ved at sætte noderne i system, også selvom de får hjælp. Ligeledes kan folk, som i højere grad tror på konspirationsteorier, måske have svært ved at fralægge sig sit systemkritiske verdenssyn, når nogen forsøger at overbevise dem om, at billeder de tror dækker over noget, ikke er sande. Resultatet peger dermed på at diffus skepsis ikke kan erstatte specifikke evner til at sondre mellem ægte og falsk information, og måske endda kan gøre mere skade end gavn.

Udover at undersøge, hvor gode folk er til at identificere ægtheden af billeder, bad vi også respondenterne forholde sig til, hvor komfortable de vil være med at dele de enkelte billeder.

På den måde kan vi undersøge, hvorvidt der er en sammenhæng mellem, hvor komfortabel danskerne er med at dele et billede, og hvorvidt de tror, at billedet er skabt med AI.

Figur 9 viser, hvordan **danskerne er mindre komfortable med at dele billeder, de tror er skabt med AI**. Det ses ved, at hældningen på linjen er negativ for alle typer af billeder. Sammenhængen er særlig stærk når det gælder billeder, der viser mennesker, hvor respondenterne er ret komfortable ved at dele billeder, de tror er ægte, men meget lidt komfortable, når de tror, billedet er AI-skabt.

Figur 9: Danskerne er mindre komfortable med at dele billeder, de tror er skabt med AI



Figuren viser ligeledes, at danskerne generelt er mere bekymrede for at dele billeder, der afbilder kriser og politik, selvom de tror, at billedet er ægte. Det kan måske skyldes, at med selv en lille frygt for, at billedet kunne være AI-skabt, så opleves der større konsekvenser ved at dele et AI-skabt billede af en skovbrand end et af et æbletræ.

Omvendt gælder det for billeder, der viser almindelige ting, som f.eks. et æbletræ eller en ko, der græsser. Her er de fleste hverken ukomfortable eller komfortable med at dele billedet med venner og familie, selvom de tror, at billedet er AI-skabt.

Betydningen af opmærksomheden rettet mod misinformation

Hvorvidt måden, vi taler om misinformation på har negative følgevirkninger, kan synes som et meget teoretisk spørgsmål. Det er dog let at forestille sig negative konsekvenser i sig selv af,

at folk går rundt og er nervøse for misinformation, og hvorvidt det påvirker demokratiske valg. En af de konsekvenser, som har været formuleret, er risikoen for en såkaldt 'løgners rente' *(liar's dividend)* (Chesney & Citron, 2019). Idéen er, at det er lettere for uærlige personer at forkaste sand information som falsk med henvisning til, at billedet, videoen eller lydklippet er skabt med AI. På den måde betaler offentligheden en rente til løgneren, som prisen for at have følehornene ude efter misinformation.

Som tankeeksperiment kan man forestille sig, hvis det var i 2024, at det for Donald Trump skadende *Access Hollywood Tape* kom ud, der kort før det amerikanske præsidentvalg i 2016 viste en Donald Trump på slap line om, hvordan man som kendt person let kan gøre sig seksuelle tilnærmelser til kvinder (Fahrenthold, 2016). Med en befolkning, der alle har hørt om *robocalls* med stemmen af Joe Biden, russiske desinformationskampagner og et hav af overbevisende AI billeder og videoer, er det ikke utænkeligt, at modargumentet om båndets autenticitet ville være lettere at fremføre i 2024 end i 2016.

Spørgeskemaets tredje tema tester derfor, hvorvidt vi bør være nervøse for, om det har negative demokratiske konsekvenser, hvis vi taler alarmistisk om misinformation. Vi følger designet fra Jungherr & Rauchfleisch (2022), der har foretaget et lignende eksperiment i USA. Det indebærer, at vi eksponerer forskellige grupper for forskellige udsagn om, hvor stort et problem misinformation er, hvorefter vi spørger respondenterne til en række udsagn, som f.eks. hvor nervøse de er for at misinformation påvirker demokratiske valg. Konkret inddeler vi tilfældigt respondenterne i tre grupper, hvor den ene gruppe får et udsagn om, at misinformation ikke rigtig er et samfundsmæssigt problem ('affejende'), den anden at hvorvidt misinformation er et samfundsmæssigt problem er usikkert ('balanceret') og en tredje at misinformation er den største demokratiske udfordring på kort sigt ('alarmistisk'). En detaljeret beskrivelse af eksperimentet kan ses i bilag 3E.

Konkret undersøger vi effekten af en fremstilling af en alarmistisk opmærksomhed på misinformation på respondenternes:

- opbakning til demokrati som styreform
- frygt for at blive beskyldt for at sprede misinformation
- frygt for at blive eksponeret for misinformation
- frygt for at AI og misinformation har negative effekter på demokratiet

Alle fire elementer er konstitueret af to-tre spørgsmål, som alle kan findes i bilag 3E.

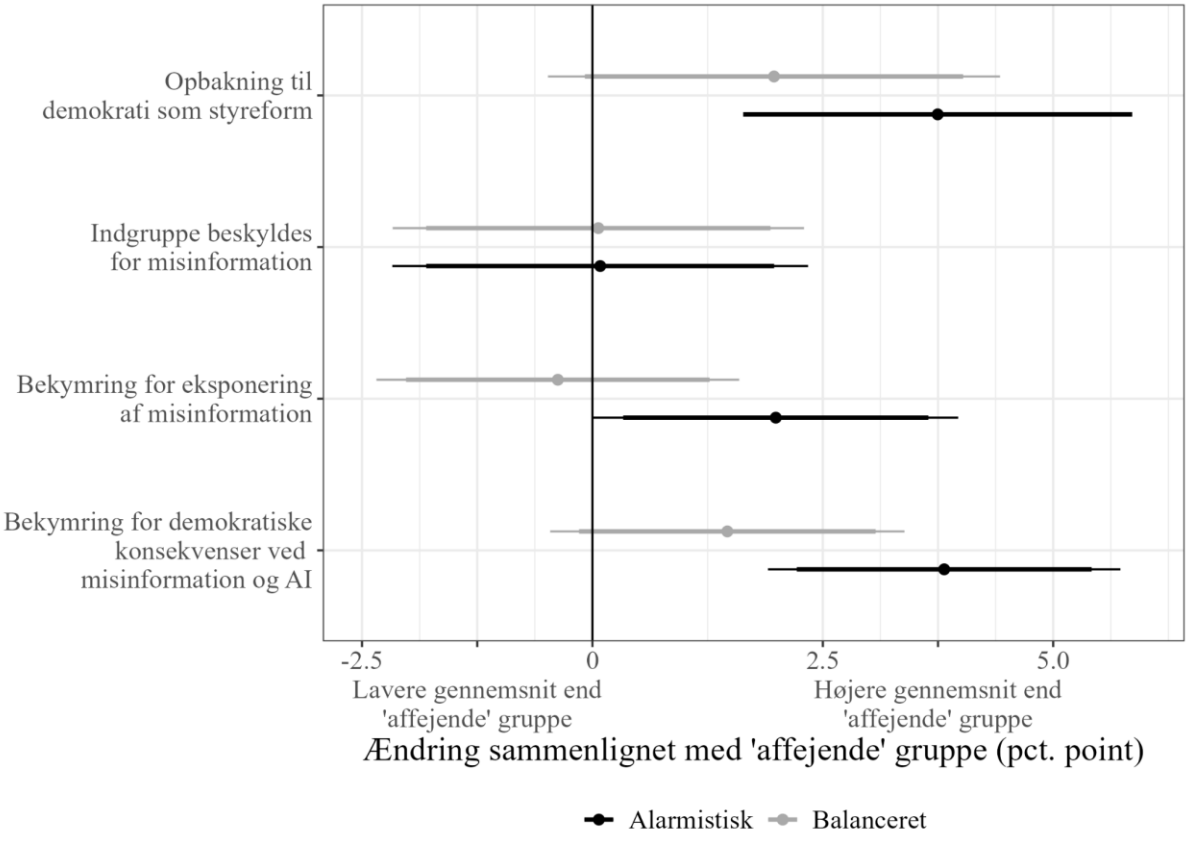
Figur 10 er et såkaldt koefficientplot. Det betyder, at figuren viser, om der er forskel mellem, hvad respondenterne i de forskellige grupper svarer, og om vi med rimelig sikkerhed kan sige, at den forskel ikke er udtryk for tilfældighed. Punkterne i figuren tolkes som forskelle fra den gruppe, der har fået at vide, at misinformation ikke er en særlig udfordring (den 'affejende' gruppe). Ser vi f.eks. på den øverste række i figuren, skal det forstås således, at den gruppe, der har modtaget den alarmistiske beskrivelse (sort prik og streg) er omtrent 3,5 procentpoint mere positive over for demokratiet som styreform end de, som har fået den affejende beskrivelse. De, der har fået den balancerede beskrivelse (grå prik og streg), er omtrent 1,6 procentpoint mere positive over for demokratiet som styreform, men da stregen overlapper 0 (sort vertikal streg) kan vi ikke med rimelig sikkerhed sige, at det ikke blot er tilfældigheder, der spiller os et puds. Her vil vi altså uddrage, at en alarmistisk beskrivelse synes at gøre folk

mere positivt stemte over for demokratiet som styreform end en affejende beskrivelse, mens det samme ikke gør sig gældende med en balanceret beskrivelse.

Figur 10 viser overordnet, hvordan **en alarmistisk udlægning af udfordringerne med misinformation har en betydning for, hvordan danskerne opfatter vigtige spørgsmål for demokratiet**.

Af negative konsekvenser bliver danskerne **mere bekymrede for at blive eksponeret for misinformation og for at misinformation og AI påvirker demokratiet og demokratiske valg**. Dette er negativt, da det kan være med til at øge risikoen for, at uærlige personer får lettere ved at snyde og slippe ud af skandaler og at folk mister tiltroen til, at hvad de ser og hører er sandt. Særligt frygten for, at misinformation kan påvirke demokratiet og demokratiske valg, kan synes farlig. I sin ekstrem kan det betyde lettere vilkår for politikere, der ikke vil acceptere valgnederlag med henvisning til manipulation fra AI og misinformation. Om end et sådan scenarie ganske vist er mindre sandsynligt i en dansk kontekst, kan sådan frygt fungere som gødning af mistillid til politikere, medier og andre samfundsmæssige aktører.

Figur 10: En alarmistisk udlægning af problemerne med misinformation øger frygten for misinformation og dens effekt, men øger også opbakning til demokratiet



Af positive konsekvenser giver danskerne udtryk for **større opbakning til demokratiet som styreform**. Det er værd at bemærke, at dette går imod vores teoretiske forventning om at finde samme mønster, som blev rapporteret i USA af Jungherr & Rauchfleisch (2022), hvorfor man skal være ekstra varsom med tolkningen heraf. Forskellen mellem vores resultater og eksperimentet i USA kan muligvis skyldes, at danskere har en større tillid til samfundets

institutioner, politikere og medier. En alarmistisk udlægning af misinformation som et problem kan muligvis fortolkes blandt danskerne som, at der bliver reageret på en samfundsmæssig udfordring. Det øger folks tillid til, at demokratiet er i stand til at identificere og adressere samfundsmæssige problematikker. Antagelsen er her, at danskerne forud for eksperimentet i sin dagligdag er blevet eksponeret for forskellige beskrivelser af udfordringer med misinformation fra journalister og politikere, hvilket gør dem skeptiske, når de får af vide, at misinformation intet problem er, og tilfredse når de hører, at nogle reagerer. Det vil dog kræve yderligere undersøgelser at fastlægge en sådan forklaring.

En anden vigtig implikation fra eksperimentet er, at der **ingen forskelle er mellem den gruppe, der bliver eksponeret for en balanceret beskrivelse af misinformation, og den der får en affejende beskrivelse af misinformation**. Denne manglende forskel kan ses positivt i det lys, at den balancerede beskrivelse nuanceret udlægger udfordringer med misinformation, uden hverken at blæse problemet op eller affeje den. Den manglende effekt betyder derfor, at en balanceret måde at kommunikere på vil give danskerne god og nuanceret information, samtidig med at man undgår negative følgevirkninger.

Vurderinger

På baggrund af resultaterne fra spørgeskemaundersøgelsen har vi følgende overordnede vurderinger:

Ældre er særligt udsatte, når det gælder risikoen for at blive snydt af indhold skabt med generativ AI. Hvor fokus i debatten ofte omhandler risikoen for børn og unge, viser vores resultater et vigtigt aspekt i også at sætte fokus på ældre, og muligt igangsætte processer med NGO'er, der arbejder med ældre om, hvordan man styrker de ældres evner i at navigere online.

Uddannelse kan måske styrke danskernes evne til at skelne mellem ægte og AI-genereret indhold. De, der er bedre til at reflektere kritisk, og de, der har prøvet kræfter med generativ AI, er generelt bedre til at skelne mellem ægte og AI-genereret indhold. Dette indikerer, at det måske er muligt at styrke danskernes evne gennem uddannelse i generativ AI, og hvad man som borger skal være opmærksom på. En sådan indsats er i overensstemmelse med den del af litteraturgennemgangen, som viste, at folk kan 'vaccineres' mod misinformation, ved at blive udsat for små mængder i kontrollerede situationer, for på den måde at blive bedre til at identificere misinformation fremadrettet. På linje med vaccinationer mod sygdomme, er det imidlertid vigtigt at tænke i, hvordan varige effekter sikres. Her er det vigtigt at udvikle folks opmærksomhed og evne til kritisk at reflektere for også at ruste befolkningen til fremtidens udviklinger. Generativ AI bliver forventeligt bedre over tid, og derfor er både et kontinuerligt og generelt fokus på teknologiforståelse afgørende, når uddannelsesinitiativer diskuteres.

Diffus skepsis er ikke gavnligt og potentielt kontraproduktivt. I forlængelse af vurderingen om betydningen af uddannelse, ligger spørgsmålet om skepsis. Her ser vi, hvordan erfaring og kritisk tænkning er positivt korreleret med evnen til at identificere ægtheden af indhold. Modsat er de med højere grad af tro på konspirationer generelt ringere. Høj grad af tro på konspirationsteorier er udtryk for en generel mistro, men sådan en mistro

er potentielt forringende for danskernes evner til at vurdere ægtheden af billeder. Det er derfor nødvendigt, at kommunikation, initiativer og uddannelse frembringer erfaring og refleksion snarere end diffus skepsis.

Det er vigtigt at udvise påpasselighed med at bruge alarmistisk sprog og huske, at danskerne godt kan forholde sig til nuanceret information. Vores resultater viser tydeligt, hvordan danskerne reagerer, når de eksponeres for et alarmistisk budskab om farerne ved misinformation. Sådanne reaktioner kan dog være negative i det omfang, at ondsindede aktører får mulighed for at snyde med henvisning til folks generelle frygt for misinformation og kunstig intelligens. Imidlertid har det ingen negative effekter at omtale farerne fra misinformation på en balanceret og nuanceret måde. Vi ved ligeledes fra bl.a. studier under Covid-19 pandemien, at transparent information fra myndighederne øger danskernes tillid til dem og sænker troen på konspirationsteorier (Petersen et al., 2021).

Danskerne har en generel høj appetit på indholdsmoderation og regulering, men der findes væsentlige forskelle mellem eksisterende og ønsket indholdsmoderation. Når danskerne blive bedt om at vurdere, hvorvidt de er tilfredse med eksisterende indholdsmoderationspolitikker, er de generelt tilfredse, om end med nogle undtagelser. Mest tydeligt gælder det spørgsmålet om, hvorvidt nøgenhed på sociale medier accepteres. Svarene giver indtryk af, at danskerne generelt ikke nyder at være underlagt et amerikansk moralkodeks for nøgenhed, om end vi ikke ud fra eksisterende data kan sige noget om, hvor vigtigt det er for danskerne. Hvad angår regulering modtog alle præsenterede forslag generelt stor opbakning.

Undersøgelsens opdrag og organisering

Som led i Medieaftalen for 2023-2026 indgår der en aftale om en årlig uvildig rapport om tech-giganternes indflydelse på demokrati, trivsel og samfundsmæssig sammenhængskraft i Danmark. Denne rapport er den første i rækken.

Rapportens formål er at danne det første grundlag i at indsamle, strukturere og producere viden om tech-giganternes indflydelse på samfundet, for på den måde at skabe et grundlag for en styrkelse af den demokratiske kontrol med tech-giganterne. Rapporterne kan over årene fokusere på forskellige og skiftende temaer afhængigt af den teknologiske udvikling, den seneste tilvejebragte viden og forskning, samt det aktuelle politiske fokus. Denne første rapport fokuserer overordnet på tech-giganternes effekter på den demokratiske samtale.

Rapporten er udarbejdet af forskere fra Digital Democracy Centre (Syddansk Universitet), SODAS (Københavns Universitet) og DATALAB (Aarhus Universitet).

Litteraturliste

Ahmad, W., Sen, A., Eesley, C., & Brynjolfsson, E. (2024). Companies inadvertently fund online misinformation despite consumer backlash. *Nature*, *630*(8015), 123-131.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, *31*(2), 211-236.

Allcott, H., Gentzkow, M., Mason, W., Wilkins, A., Barberá, P., Brown, T., ... & Tucker, J. A. (2024). The effects of Facebook and Instagram on the 2020 election: A deactivation experiment. *Proceedings of the National Academy of Sciences*, *121*(21), e2321584121.

Allen, J., Watts, D. J., & Rand, D. G. (2024). Quantifying the impact of misinformation and vaccine-skeptical content on Facebook. *Science*, *384*(6699), eadk3451.

Akram, M., Nasar, A., & Arshad-Ayaz, A. (2022). A bibliometric analysis of disinformation through social media. *Online Journal of Communication and Media Technologies*, *12*(4), e202242.

Amnesty International (2022). Myanmar: Facebook's systems promoted violence against rohingya; Meta owes reparations. *Amnesty.org*. https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/

Aral, S. (2021). *The hype machine: How social media disrupts our elections, our economy, and our health--and how we must adapt*. *Crown Currency.*

Altay, S., Berriche, M., & Acerbi, A. (2023). Misinformation on misinformation: Conceptual and methodological challenges. *Social media+ Society*, *9(1)*, 20563051221150412.

Altay, S., Nielsen, R. K., & Fletcher, R. (2022). Quantifying the "infodemic": People turned to trustworthy news outlets during the 2020 coronavirus pandemic. *Journal of Quantitative Description: Digital Media*, *2*.

Baptista, J. P., & Gradim, A. (2020). Understanding fake news consumption: A review. *Social Sciences*, *9*(10), 185.

Blanchard, L., Conway-Moore, K., Aguiar, A., Önal, F., Rutter, H., Helleve, A., ... & Knai, C. (2023). Associations between social media, adolescent mental health, and diet: A systematic review. *Obesity Reviews*, *24*, e13631.

Bor, A., & Petersen, M. B. (2022). The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. *American Political Science Review*, *116*(1), 1-18.

Bryanov, K., & Vziatysheva, V. (2021). Determinants of individuals' belief in fake news: A scoping review determinants of belief in fake news. *PLoS one*, *16*(6), e0253717.

Budak, C., Nyhan, B., Rothschild, D. M., Thorson, E., & Watts, D. J. (2024). Misunderstanding the harms of online misinformation. *Nature*, *630*(8015), 45-53.

Børns Vilkår (2024). Børns liv med sociale medier: Hvordan forholder børn sig til videoindhold, influencere og AI-chatbots. *Børns Vilkår*. https://bornsvilkar.dk/wp-content/uploads/2024/05/Boern-og-unges-liv-paa-sociale-medier_enkeltsidet.pdf

Center for Sociale Medier, Tech og Demokrati (2023). *Danskernes holdning til den demokratiske samtale på online platforme*. Center for Sociale Medier, Tech of Demokrati. https://slks.dk/fileadmin/user_upload/SLKS/Omraader/Medier/Tech-center/Undersoegelse_-_Danskernes_holdning_til_den_demokratiske_samtale_paa_online_platforme.pdf

Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, *107*, 1753.

Danmarks Statistik (2023). *Danmark bruger sociale medier mest i EU*. Danmarks Statistik. https://www.dst.dk/da/Statistik/nyheder-analyser-publ/nyt/NytHtml?cid=46771

Danske Medier Research (2024). *Annoncemarked*. Danske Medier. https://danskemedier.dk/branchetal-statistik/marked/

Darius, P. (2024). Researcher Data Access Under the DSA: Lessons from TikTok's API Issues During the 2024 European Elections. *Tech Policy.press*. https://www.techpolicy.press/-researcher-data-access-under-the-dsa-lessons-from-tiktoks-api-issues-during-the-2024-european-elections/

Dommett, K. (2023). The inter-institutional impact of digital platform companies on democracy: A case study of the UK media's digital campaigning coverage. *new media & society*, *25*(10), 2763-2780.

DR Analyse (2024). *Medieudviklingen 2023*. DR https://www.dr.dk/om-dr/fakta-om-dr/medieforskning/medieudviklingen/2023

Duffy, Clare (2023). *Elon Musk's X is encouraging users to follow conspiracy theorist Alex Jones after reinstating his account*. CNN https://edition.cnn.com/2023/12/11/tech/elon-musk-x-promoting-alex-jones-after-reinstating-his-account/index.html

Eady, G., Bonneau, R., Tucker, J. A., & Nagler, J. (under udgivelse). News sharing on social media: Mapping the ideology of news media content, citizens, and politicians. *Political Analysis*.

Ecker, U., Roozenbeek, J., van der Linden, S., Tay, L. Q., Cook, J., Oreskes, N., & Lewandowsky, S. (2024). Misinformation poses a bigger threat to democracy than you might think. *Nature*, *630*(8015), 29-32.

Eckles (2023). *thefacebook and mental health trends: Harvard and Suffolk County Community College* https://statmodeling.stat.columbia.edu/2023/08/22/thefacebook-and-mental-health-trends-harvard-and-suffolk-community-college/

Ergün, N., Özkan, Z., & Griffiths, M. D. (2023). Social media addiction and poor mental health: examining the mediating roles of internet addiction and phubbing. *Psychological reports*, 00332941231166609.

Erhvervsministeriet (2024). Grænser for tech-giganternes udvikling og anvendelse af kunstig intelligens: Delafrapportering 2 fra regeringens ekspertgruppe om tech-giganter. *Erhvervsministeriet*. https://www.em.dk/Media/638460136860682710/web_Tech-ekspertgruppens%20afrapportering%20vedr.%20AI%20MF_v01_140324.pdf

Fahrenthold, D. (2016). Trump recorded having extremely lewd conversation about women in 2005. *Washington Post*. https://www.washingtonpost.com/politics/trump-recorded-having-extremely-lewd-conversation-about-women-in-2005/2016/10/07/3b9ce776-8cb4-11e6-bf8a-3d26847eeed4_story.html

Ferguson, C. J. (2024). Do social media experiments prove a link with mental health: A methodological and meta-analytic review. *Psychology of Popular Media*.

Garrett, R. K., & Stroud, N. J. (2014). Partisan paths to exposure diversity: Differences in pro-and counterattitudinal news consumption. *Journal of Communication*, *64*(4), 680-701.

Golovchenko, Y., Buntain, C., Eady, G., Brown, M. A., & Tucker, J. A. (2020). Cross-platform state propaganda: Russian trolls on twitter and YouTube during the 2016 US Presidential Election. *The International Journal of Press/Politics*, *25*(3), 357-389.

Golovchenko, Y. (2022). Fighting propaganda with censorship: A study of the Ukrainian ban on Russian social media. *The Journal of Politics*, *84*(2), 639-654.

Guess, A., Nyhan, B., Lyons, B., & Reifler, J. (2018). Avoiding the echo chamber about echo chambers. *Knight Foundation*, *2*(1), 1-25.

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., ... & Tucker, J. A. (2023a). Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science*, *381*(6656), 404-408.

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., ... & Tucker, J. A. (2023b). How do social media feed algorithms affect attitudes and behavior in an election campaign?. *Science*, *381*(6656), 398-404.

Greenspan, R. L., & Loftus, E. F. (2021). Pandemics and infodemics: Research on the effects of misinformation on memory. *Human Behavior and Emerging Technologies*, *3*(1), 8-12.

Haidt, J. (2024). *The anxious generation: How the great rewiring of childhood is causing an epidemic of mental illness*. Random House.

Hove, M. F. (2024). Hvem målretter dig og virker det? Politikernes brug af annoncer på sociale medier. *Magtudredningen 2.0*. https://ps.au.dk/fileadmin/Statskundskab/Billeder/Forskning/Forskningsprojekter/Magtudredning/Essays/Tema15/Mads_Fuglsang_Hove.pdf

Hove, M. F., Hobolt, S. B., van Dalen, A., & de Vreese, C. H. (2024). Partiernes brug af online politisk microtargeting. In *Partiledernes kamp om midten: Folketingsvalg 2022*. Djøf Forlag.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, *22*(1), 129-146.

Jungherr, A., & Rauchfleisch, A. (2022). Negative downstream effects of disinformation discourse: evidence from the US. *Preprint at SocArXiv https://doi. org/10.31235/osf. io/a3rzm*.

Khalaf, A. M., Alubied, A. A., Khalaf, A. M., & Rifaey, A. A. (2023). The impact of social media on the mental health of adolescents and young adults: a systematic review. *Cureus*, *15*(8).

Kim, B., Xiong, A., Lee, D., & Han, K. (2021). A systematic review on fake news research through the lens of news creation and consumption: Research efforts, challenges, and future directions. *PloS one*, *16*(12),

King, G., Pan, J., & Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. *American political science Review*, *107*(2), 326-343.

Kulturministeriet (2021). Annonceomsætning 2021. *Kulturministeriet, Mediernes Udvikling*. https://kum.dk/kulturomraader/medier/mediernes-udvikling/publikationer/annonceomsaetning-2021

Leerssen, P., Dobber, T., Helberger, N., & de Vreese, C. (2023). News from the ad archive: How journalists use the Facebook Ad Library to hold online advertising accountable. *Information, communication & society*, *26*(7), 1381-1400.

Lu, Y., & Lee, J. K. (2019). Partisan information sources and affective polarization: Panel analysis of the mediating role of anger and fear. *Journalism & Mass Communication Quarterly*, *96*(3), 767-783.

Nisgaard, Allan (2024). Ekspert om falske kendis annoncer på Facebook: Alle kan blive ramt af den. DR https://www.dr.dk/nyheder/viden/teknologi/ekspert-om-falske-kendis-annoncer-paa-facebook-alle-kan-blive-ramt-af-den

Nyhan, B., Settle, J., Thorson, E., Wojcieszak, M., Barberá, P., Chen, A. Y., ... & Tucker, J. A. (2023). Like-minded sources on Facebook are prevalent but not polarizing. *Nature*, *620*(7972), 137-144.

Hancock, J., Liu, S. X., Luo, M., & Mieczkowski, H. (2022). Psychological well-being and social media use: A meta-analysis of associations between social media use and depression, anxiety, loneliness, eudaimonic, hedonic and social well-being. *Anxiety, Loneliness, Eudaimonic, Hedonic and Social Well-Being (March 9, 2022)*.

Hasell, A., & Weeks, B. E. (2016). Partisan provocation: The role of partisan news use and emotional responses in political information sharing in social media. *Human Communication Research*, *42*(4), 641-661.

Kaushik, D. (2024). Policy Responses To Fake News On Social Media Platforms: A Law And Economics Analysis. *Statute Law Review*, *45*(1), hmae013.

Levendusky, M. S. (2013). Why do partisan media polarize viewers?. *American Journal of Political Science*, *57*(3), 611-623.

Martel, C., & Rand, D. G. (2023). Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology*, 101710.

McCabe, S. D., Ferrari, D., Green, J., Lazer, D. M., & Esterling, K. M. (2024). Post-January 6th deplatforming reduced the reach of misinformation on Twitter. *Nature*, *630*(8015), 132-140.

Meta (2024). About social issues. *Meta Business Help Center.* https://www.facebook.com/business/help/214754279118974?id=288762101909005

Miller, J., Mills, K. L., Vuorre, M., Orben, A., & Przybylski, A. K. (2023). Impact of digital screen media activity on functional brain organization in late childhood: Evidence from the ABCD study. *cortex*, *169*, 290-308.

Morrow, G., Swire-Thompson, B., Polny, J. M., Kopec, M., & Wihbey, J. P. (2022). The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology*, *73*(10), 1365-1386.

Murray, Conor (2021). TikTok algorithm error sparks allegations of racial bias. *NBC News*. https://www.nbcnews.com/news/us-news/tiktok-algorithm-prevents-user-declaring-support-black-lives-matter-n1273413

Nanz, A., & Matthes, J. (2022). Democratic consequences of incidental exposure to political information: A meta-analysis. *Journal of Communication*, *72*(3), 345-373.

Nasery, M., Turel, O., & Yuan, Y. (2023). Combating fake news on social media: a framework, review, and future opportunities. *Communications of the Association for Information Systems*, *53*(1), 833-876.

Nyhan, B., Settle, J., Thorson, E., Wojcieszak, M., Barberá, P., Chen, A. Y., ... & Tucker, J. A. (2023). Like-minded sources on Facebook are prevalent but not polarizing. *Nature*, *620*(7972), 137-144.

Newman, N., Fletcher, R., Robertson, C., Arguedas, A. & Nielsen, R. (2024). Reuters Institute, Digital News Report 2024 https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2024-06/DNR%202024%20Final%20lo-res-compressed.pdf

Odgers, C. L., & Jensen, M. R. (2020). Annual research review: Adolescent mental health in the digital age: Facts, fears, and future directions. *Journal of Child Psychology and Psychiatry*, *61*(3), 336-348.

Odgers, C. L. (2024). The great rewiring: is social media really behind an epidemic of teenage mental illness?. *Nature*, 628(8006), 29-30.

Orben, A. (2020). Teenagers, screens and social media: a narrative review of reviews and key studies. *Social Psychiatry and Psychiatric Epidemiology*, *55*(4), 407-414.

Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature human behaviour*, *3*(2), 173-182.

Oversight Board (2023). *Oversight Board Overturns Meta's Original Decisions in the "Gender Identity and Nudity" Cases.* Meta's Oversight Board. https://www.oversightboard.com/news/1214820616135890-oversight-board-overturns-meta-s-original-decisions-in-the-gender-identity-and-nudity-cases/

Pariser, E. (2011). *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in cognitive sciences*, *25*(5), 388-402.

Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature communications*, *13*(1), 2333.

Petersen, M. B., Bor, A., Jørgensen, F., & Lindholt, M. F. (2021). Transparent communication about negative features of COVID-19 vaccines decreases acceptance but increases trust. *Proceedings of the National Academy of Sciences*, *118*(29), e2024597118.

Pew Research Center (2024). *About half of TikTok users under 30 say they use it to keep up with politics, news.* Pew Research Center. https://www.pewresearch.org/short-reads/2024/08/20/about-half-of-tiktok-users-under-30-say-they-use-it-to-keep-up-with-politics-news/

Quétier-Parent, S., Lamotte, D., Gallard, M. (2023). *Elections & social media: the battle against disinformation and trust issues.* Ipsos. https://www.ipsos.com/en/elections-social-media-battle-against-disinformation-and-trust-issues

Radtke, T., Apel, T., Schenkel, K., Keller, J., & von Lindern, E. (2022). Digital detox: An effective solution in the smartphone era? A systematic literature review. *Mobile Media & Communication*, *10*(2), 190-215.

Reeves, B., Robinson, T., & Ram, N. (2020). Time for the human screenome project. *Nature*, *577*(7790), 314-317.

Sanders, A. K., & Jones, R. L. (2018). Clicks at Any Cost: Why Regulation Won't Upend the Economics of Fake News. *Bus. Entrepreneurship & Tax L. Rev.*, *2*, 339.

Shahzad, K., Khan, S. A., Iqbal, A., Shabbir, O., & Latif, M. (2023). Determinants of fake news diffusion on social media: A systematic literature review. *Global Knowledge, Memory and Communication*.

Stevenson, A. (2018). Facebook Admits It Was Used to Incite Violence in Myanmar. *The New York Times.* https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html

Sunstein, C. R. (2001). *Republic. com*. Princeton University Press.

Thadani, T. (2024). Elon Musk's X accused of bias after pro-Harris accounts labeled as 'spam'. *The Washington Post* https://www.washingtonpost.com/technology/2024/08/07/musk-x-harris-bias/

Thorp, H. H. (2024). Unsettled science on social media. *Science*.

Twenge, J. M., Haidt, J., Joiner, T. E., & Campbell, W. K. (2020). Underestimating digital media harm. *Nature Human Behaviour*, *4*(4), 346-348.

Uldall, Rosa (2024). DF deler deepfake video af statsministeren: tendensen kan være undergravende. *DR.* https://www.dr.dk/nyheder/indland/df-deler-deepfake-video-af-statsministeren-tendensen-kan-vaere-undergravende

Valkenburg, P. M., Meier, A., & Beyens, I. (2022). Social media use and its impact on adolescent mental health: An umbrella review of the evidence. *Current Opinion in Psychology*, *44*, 58-68.

Van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151.

Walter, N., & Tukachinsky, R. (2020). A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it?. *Communication research*, *47*(2), 155-177.

Walter, N., Brooks, J. J., Saucier, C. J., & Suresh, S. (2021). Evaluating the impact of attempts to correct health misinformation on social media: A meta-analysis. *Health communication*, *36*(13), 1776-1784.

Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication monographs*, *85*(3), 423-441.

Weigle, P. E., & Shafi, R. M. (2024). Social media and youth mental health. *Current psychiatry reports*, *26*(1), 1-8.

Weiss, G. (2024). We now know just how much money Elon Musk's X made after his takeover — and it's a lot less than before his purchase, *Business Insider* https://www.businessinsider.com/x-revenues-plunged-months-after-elon-musk-took-over-report-2024-6

World Economic Forum (2024). Global Risks 2024: Disinformation Tops Global Risks 2024 as Envrionmental Threats Intensify. *World Economic Forum* https://www.weforum.org/press/2024/01/global-risks-report-2024-press-release/

Wu, Y., Ngai, E. W., Wu, P., & Wu, C. (2022). Fake news on the internet: a literature review, synthesis and directions for future research. *Internet Research*, *32*(5), 1662-1699.

Yang, Y., McCabe, S., & Hindman, M. (2024). Does Russian Propaganda Lead or Follow? Topic Coverage, User Engagement, and RT and Sputnik's Agenda Influence on US Media. *The International Journal of Press/Politics*, 19401612241271074.

Zuleta, L. & Burkal, R. (2017). Hadefulde ytringer i den offentlige online debat. *Institut for Menneskerettigheder.*

Bilag

Bilag 1A: Indsamling af aktivitets- og annoncedata fra Facebook

Data omhandlende antallet af dagligt aktive brugere er indsamlet fra Facebook Marketing API. API'en er oprettet og vedligeholdt af Facebook selv, og giver godkendte udviklere adgang til data om, hvor mange brugere, der er aktive på Facebook inden for forskellige kategorier. Således har vi foretaget API kald, hvor vi har spurgt efter information om, hvor mange aktive brugere, der har været på Facebook i løbet af forskellige ugedage i september 2024, hvoraf vi har taget en gennemsnit for at undgå effekter af forskellige aktivitet i løbet af ugen. Den information, der sendes retur inkluderer et estimat for antallet af aktive brugere inklusiv et nedre og et øvre estimat, hvilket ligeledes er visualiseret i figur 1 og 2.

Til figur 1 har vi indhentet data om antallet af aktive brugere for hvert land, mens figur 2 alene inkluderer brugere i Danmark. I figur 2 har vi foretaget API kald for en række aldersgrupper, samt antallet af brugere inden for hver af aldersgrupperne, der er mænd og kvinder.

Data omhandlende annonceomsætning af annoncer på Facebook og Instagram er indsamlet fra Meta Ad Library API (annoncebibliotek). De indhentede data inkluderer udelukkende annoncer om samfundsmæssige forhold, da det er den eneste type af annoncer, der er mulige at tilgå længere end 12 måneder bagud i tid, ligesom de på nuværende tidspunkt er de eneste annoncer, hvor der er data på omkostninger i kroner ved de enkelte annoncer.

Konkret har vi hentet information om samtlige annoncer om samfundsforhold fra perioden 1. januar 2020 til 1. juli 2024 og estimeret den månedlige omsætning ved at udregne, hvor stor en andel en given annonces budget er placeret i de enkelte måneder, givet at en annonces forbrug er konstant. F.eks. vil en annonce, der har været indrykket 1. januar 2020 til 15. februar 2020 med et samlet forbrug på 150 kroner blive opgjort således, at annoncen har kostet 100 kroner i januar og 50 kroner i februar. Da der ikke er adgang til estimater for det samlede annoncepres for en given dag eller måned er det ikke muligt at tage højde for, at nogle dage eller måneder er dyrere end andre. Dog er udfordringerne forbundet med dette mindre, da vi afrapporterer udviklingen i figur 3 med lokal regression. Derudover er udviklingen i figur 3 ligeledes visualiseret med usikkerhedsestimater, der tager udgangspunkt i, at forbruget for en given annonce afrapporteres i et spænd, f.eks. mellem 100 og 200 kroner. Udviklingen er derfor udregnet ud fra spændets gennemsnit, mens den visualiserede usikkerhed angiver henholdsvis det nedre og øvre estimat.

Vi har i rapporten valgt at fokusere på Facebook og Instagram, samt udelukkende annoncer om samfundsforhold, da det er disse platforme og disse annoncer, der har været åbenhed om længst og hvor der forventeligt er den højeste datakvalitet at finde. Fremgangsmåden kan imidlertid på sigt udvides til både at omfatte flere platforme og andre typer af annoncer. Som led i EU-forordningen om digitale tjenester (DSA) er alle store online platforme (VLOP) nødt til at oprette offentligt tilgængelige annoncebiblioteker indeholdende samtlige annoncer, der er aktive eller har været aktive på deres platforme det seneste år. For disse annoncebiblioteker er det ligeledes et krav, at forskere skal kunne få adgang til en API, således data kan trækkes direkte fra plaformen og struktureres. Adgangene er dog fortsat nye og får fortsat kritik for

kvaliteten af data (Darius, 2024). Det vil derfor – forhåbenligtvis – på sigt være muligt at anvende sådanne adgang til at estimere den samlede annonceomsætning på de store online platforme.

Bilag 2A: Identifikation og systematisering af forskningslitteratur

Identifikationen af forskningslitteratur er sket i tre faser. For det første har vi haft søgt efter *review* artikler og metastudier på *Web of Science*[8]. For det andet har vi inkluderet empiriske artikler i store tidsskrifter, der er kommet ud for nyligt, og som derfor ikke har været inkluderet i de identificerede *review* artikler. For det tredje har vi haft delt de enkelte dele af litteraturgennemgangen (misinformation, polarisering og trivsel) med andre forskere på danske universiteter, der forsker i felterne, for input til, at de centrale artikler på feltet er inkluderet.

Litteratursøgningen har fokuseret på at indfange de vigtigste debatter og empiriske fund inden for forskningslitteraturen om misinformation, politisk polarisering og trivsel i relation til sociale medier. Derfor har det været en nødvendig betingelse for, at studier er blevet inkluderet i litteraturgennemgangen, at det er forholdet mellem sociale medier og det pågældende forskningsfelt, artiklen har omhandlet. Det betyder blandt andet, at artikler, der primært fokuserer på spørgsmålet om misinformation i forbindelse med Covid-19, og kun sekundært på sociale mediers betydning herfor, ikke inkluderes. Ligeledes vælger vi af afgrænsnignsmæssige årsager ikke at inkludere den del af litteraturen, som specifikt fokuserer på, hvordan man ved hjælp af statistiske modeller og lignende er i stand til at idenficiere misinformation, da dette mere handler om computationelle metoder mere end forholdet mellem sociale medier og misinformation. Tabellen herunder viser en oversigt over de forskellige skridt, der har været taget i forbindelse med første fase af identifikation af litteratur om misinformation og hvor mange artikler, der har været identificeret.

Skridt	**Antal artikler**
Søgning efter artikler med søgeordene (i titel eller abstract): AI OR generative AI OR generative OR big tech OR platforms OR social media OR Facebook OR Instagram OR YouTube OR Google OR Twitter OR Pinterest OR LinkedIn OR Snapchat OR TikTok OR Bing AND misinformation OR disinformation OR conspiracies OR conspiracy theory OR conspiracy OR fake news OR false information OR hoax OR fraud OR information disorder OR malinformation OR rumors OR propaganda OR deepfake OR infodemic OR information warfare	14.834
Anvendelse af filtre + Sprog (engelsk) og årstal (2004 - 2024)	14.834

[8] Grundet rapportens fokus på misinformation har denne første fase af identifikation af litteratur kun blevet foretaget for forskningslitteraturen om misinformation.

+ Forskningsfelt (kommunikationsvidenskab og statskundskab) + Dokument type (review)	4.052 158
Abstract *screening*	**Relevante: 36** Ikke rette forskningsfeltd: 32 Ikke empiriske eller ikke reviews: 20 Tekniske (*fake news detection*): 39 Ikke primær fokus på misinformation: 16 Studier primær fokuseret på Covid-19: 11 Afgrænset til specifikke lande: 4
Review artikler inkluderet efter læsning	17

Bilag 3A: Holdninger til indholdsmoderation på Facebook

Præcise formuleringer i spørgeskema og *community standards* til baggrund for modereringsbatteri. Formulering i *community standards* er markeret med kursiv.

Trusler om vold der kan føre til alvorlig menneskelig eller materiel skade
Trusler om vold, der kan føre til alvorlig skade (mindre grov vold). Vi fjerner sådanne trusler mod offentlige personer og grupper, der ikke er baseret på beskyttede karakteristika, hvis de er troværdige, og vi fjerner dem mod alle andre mål (herunder grupper baseret på beskyttede karakteristika) uanset troværdighed

Billeder, der viser meget tynde mennesker på en måde, der er associeret med spiseforstyrrelser
Indhold, der fokuserer på afbildning af ribben, kraveben, lårmellemrum, hofter, indbuet mave eller fremstående rygsøjle eller skulderblade, når de deles sammen med udtryk, der er associeret med spiseforstyrrelser.

Billeder af topløse badende kvinder på en strand, som er lagt op med samtykke fra kvinderne
Billeder af virkelige nøgne voksne, hvis det afbilder: Utildækkede kvindelige brystvorter undtagen i forbindelse med amning, fødsel og øjeblikkene efter en fødsel, medicinsk eller sundhedsmæssig kontekst (f.eks. efter fjernelse af bryster, øget kendskab til brystkræft eller kønsskifteoperationer) eller som en protesthandling.

Indhold, der omtaler religiøse grupper som eksempelvis dumme eller idioter
Indhold, der er rettet mod en person eller gruppe af personer på baggrund af deres beskyttede karakteristika (race, etnicitet, national oprindelse, religiøst tilhørsforhold, kaste, seksuel orientering, køn, kønsidentitet og alvorlig sygdom), med: Mentale mangler er defineret som handlende om: Intellektuelle evner, herunder, men ikke begrænset til, dum, stupid, idioter. Uddannelse, herunder, men ikke begrænset til, analfabet, uvidende. Mental sundhed, herunder, men ikke begrænset til, mentalt syg, retarderet, skør, vanvittig.

Oplysninger, der er i modstrid med den sundhedsfaglige evidens, f.eks. udsagn som: “Et stigende antal vaccinationer forklarer, hvorfor så mange børn har autisme i dag”
Forkerte oplysninger om vacciner. Vi fjerner forkerte oplysninger primært om vacciner, hvis offentlige sundhedsmyndigheder vurderer, at oplysningerne er falske og formentlig kan bidrage direkte til vaccinenægtelse. De omfatter: Vacciner forårsager autisme (f.eks. "Et stigende antal vaccinationer forklarer, hvorfor så mange børn har autisme i dag").

Indhold der krænker ophavsrettigheder
Efter modtagelse af en anmeldelse fra en rettighedsindehaver eller en autoriseret repræsentant fjerner eller begrænser vi indhold, der: Krænker ophavsrettigheder. Krænker varemærkerettigheder.

Principper for udvælgelse af temaer i modereringsbatteri
Meta skelner mellem "Vi fjerner" og "Vi fjerner når vi har lidt mere kontekst". Vi anvender udelukkende politikker fra førstnævnte kategori for at undgå tvetydeligheder.

Meta har seks kategorier i deres "community standards" der beskriver forskellige typer af indhold der modereres. 1) Violence and criminal behavior, 2) Satefy, 3) Objectionable content, 4) Integrity and authenticity, 5) Respecting intellectual property, 6) Content-related requests and decisions. Vi har eksempler med på alle undtagen punkt 6, da denne ikke inkluderer indhold, der fjernes uden yderligere kontekst.

Der er stor forskel variation i, hvor voldsomt og indgribende det, der modereres er. Vi blander så vi får variation i typen af indhold, men undgår ekstremer vi forventer alle vil være enige i, f.eks. trusler til enkeltpersoner om at slå dem ihjel..

Bilag 3B: Holdninger til reguleringsanbefalinger

De for respondenterne præsenterede reguleringer er taget fra regeringens ekspertgruppe om tech-giganter, der til erhvervsministeriet har afrapportert anbefalinger. Disse anbefalinger fordeler sig på fire temaer, hvoraf vi har udvalgt en fra hver af temaerne om *tech-giganternes medansvar for informationstroværdighed på deres platforme* og *regulering af tech-giganters uaftoriserede brug af ophavsretlig materiale*. Derudover har vi udvalgt to spørgsmål fra temaet om *beskyttelse af børn og unge mod skadelig anvendelse og udvikling af kunstig intelligens på tech-giganternes tjenester*. Vi har ikke udvalgt nogen fra temaet om *tech-giganters markedsdominans inden for udvikling af kunstig intelligens* da vi vurderede anbefalingerne for generelle til at kunne indfange holdninger til det i et spørgeskema. Respondenter er blevet stillet følgende spørgsmål:

Hvor enig eller uenig er du i at sociale medier bør …
- deklarere indhold genereret af kunstig intelligens
- tage højde for børn og unge som særligt sårbare grupper, når det gælder manipulation og afhængighed
- dokumentere at de ikke bryder ophavsretten
- undlade at kræve betaling for at brugerne kan fravælge manipulerende design

Bilag 3C: Design af AI-billede genkendelse eksperiment

Eksperimentet er designet med henblik på at måle, hvor gode individer er til at adskille, hvilke billeder, som er ægte, og hvilke billeder, der er skabt med generativ AI. For at undgå, at respondenter gætter rigtigt ved et tilfælde, bliver alle respondenter eksponeret for syv billeder. Randomisering for, hvorvidt billedet er ægte eller skabt med generativ AI varieres for hvert billede.

Hver respondent bliver eksponeret for et billede, som enten viser et ægte billede eller AI-genereret billede af samme situation. For at undgå utilsigtede stimulus effekter ved, at et enkelt billede er let identificerbart eller at folk er bedre til at identificere AI i bestemte situationer (f.eks. mennesker), benytter vi os af stimulus *sampling*. Det er en logik, hvor stimulus udtrækkes tilfældigt fra en pulje af stimuli.

Puljen af billeder består af i alt 52 billeder, hvoraf halvdelen er AI-skabt og halvdelen af ægte, da der for hver situation er både et ægte og AI-skabt billede. Yderligere fordeler billederne sig på fire kategorier: almindelig, mennesker, krise og politik.

"Almindelige" billeder inkluderer billeder som f.eks. et æbletræ og det kinesiske *bullet train*. "Mennesker" inkluderer billeder af f.eks. Jeff Bezos og en ældre kvinde, som modtager en vaccination. "Krise" inkluderer billeder af krisesituationer, f.eks. et flystyrt i Mogadishu og skovbrand i Californien. "Politik" inkluderer billeder, der kan lede tanken hen på politiske situationer og diskussioner, f.eks. Brandenburger Tor oplyst i ukrainske farver og Za'atri flygtningelejren for syriske flygtninge. Det er værd at bemærke, at der for nogle af billederne optræder mennesker i andre kategorier end "mennesker", men hvilke er billeder hvor personerne ikke er i fokus.

Ægte billeder er fundet på wikimedia.org, hvor udelukkende billeder uden *copyright* er blevet valgt. AI-billederne er skabt med det gratis og offentligt tilgængelig openart.ai. Hvert billede har taget 5-15 minutter at skabe ved at taste et par forskellige *prompts*, hvorefter det bedste af de genererede billeder er blevet valgt til at indgå i eksperimentet. Måden AI-billederne er skabt på er tiltænkt at efterligne den situation, hvor mindre aktører, f.eks. dem interesserede i at lokke social medie brugere ind på deres hjemmeside, skaber og deler billeder. Derfor vil billedekvaliteten sandsynligvis være større, havde det været del af en større statslig desinformationskampagne. Det gør imidlertid kun vores test lettere, hvorfor resultatet hvor respondenterne klarer sig omtrent på niveau med tilfældighed forventes at være en øvre bare sammenlignet med større og mere sofistikerede desinformationskampagner.

Som et eksempel på et ægte og AI-skabt billede er følgende billeder, der skal vise folk observere en solformørkelse på *The Mall* i Washington D.C. Billedet til venstre er AI-skabt og billedet til højre ægte.

Inden respondenten eksponeres og bedes forholde sig til, hvorvidt de tror billedet, de har set er skabt med AI og hvor komfortable de vil være med at dele det, får de en introduktionstekst. Teksten lyder således:

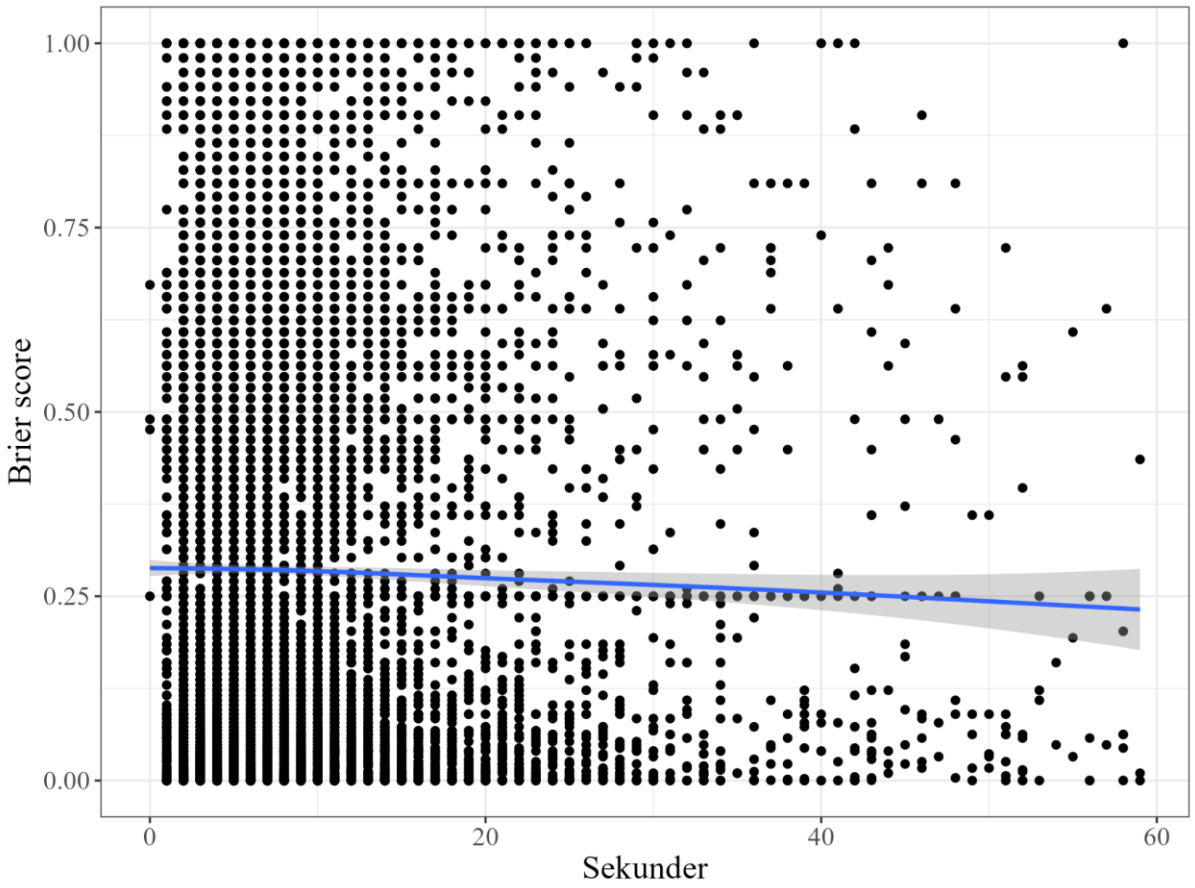
Du bliver nu præsenteret for en række billeder, der alle viser begivenheder eller situationer, som er sket i virkeligheden. Nogle af billederne er faktiske billeder, mens andre er skabt med generativ kunstig intelligens ("AI"). Du bliver bedt om at give dit umiddelbare indtryk af, om du tror at billedet er skabt med kunstig intelligens, samt hvor komfortabel du vil være med at dele billedet til familie og venner.

Som det fremgår af introduktionstesten *primer* vi respondenterne ved at fortælle dem, at nogle af de billeder de møder er skabt med AI. Det er imidlertid svært at forestille sig et design, hvor respondenterne bliver bedt om at vurdere om et billede er skabt med AI uden at *prime* dem på den ene eller anden måde. Implikationen af dette designvalg er, at respondenterne burde være *bedre* til at identificere ægtheden af billeder, nu hvor de er opmærksomme på det. Da resultatet af eksperimentet var, at respondenterne i gennemsnit klarede sig omtent på niveau med tilfældighed, ville vi ikke forvente de var bedre, havde de *scrollet* i gennem deres Instagram *feed* uden at overveje, hvor vidt det de hurtigt skyder forbi er skabt med AI.

En anden forhåbning ved introduktionsteksten var, at respondenterne ikke ville bruge for lang tid på at kigge på billederne. Dette skyldes, at vi prøver at efterligne en situation, hvor man som på sociale medier bliver eksponeret for en række billeder, og ikke en quiz i om man kan identificere AI-billeder. For at tjekke, hvorvidt respondenterne fulgte opfordring om at give deres umiddelbare indtryk, tjekkede vi, hvor mange sekunder hver respondent brugte på billede-siden af spørgeskemaet. Hver respondent brugte i gennemsnit 7 sekunder (median) på at kigge på billederne. For at teste robustheden af resultaterne har vi kørt analysen kun med billede-opgaver, hvor respondenten brugte 20 eller færre sekunder på siden. Vi finder ingen forskel i resultaterne og beholder derfor alle respondenter i hovedanalysen.

Det betyder imidlertid ikke, at respondenterne er lige ringe til at vurdere ægtheden af billeder uanset, hvor lang tid de bruger på at se på billeder. Som figur 3C1 viser, er der en svag negativ sammenhæng (0.01 højere brier score hver 10 sekund) mellem hvor mange sekunder, der har været brugt på at se på et billede, og hvor god vurderingen af billedets ægthed er. Sammenhængen beror på en *pooled* regression, da der er så lille variation inden for respondenter i forhold til, hvor mange sekunder de bruger på at kigge på billederne, at det ikke er muligt at anvende variationen inden for enheder meningsfuldt.

Figur 3C1: Sammenhæng mellem sekunder brugt på opgave og brier score



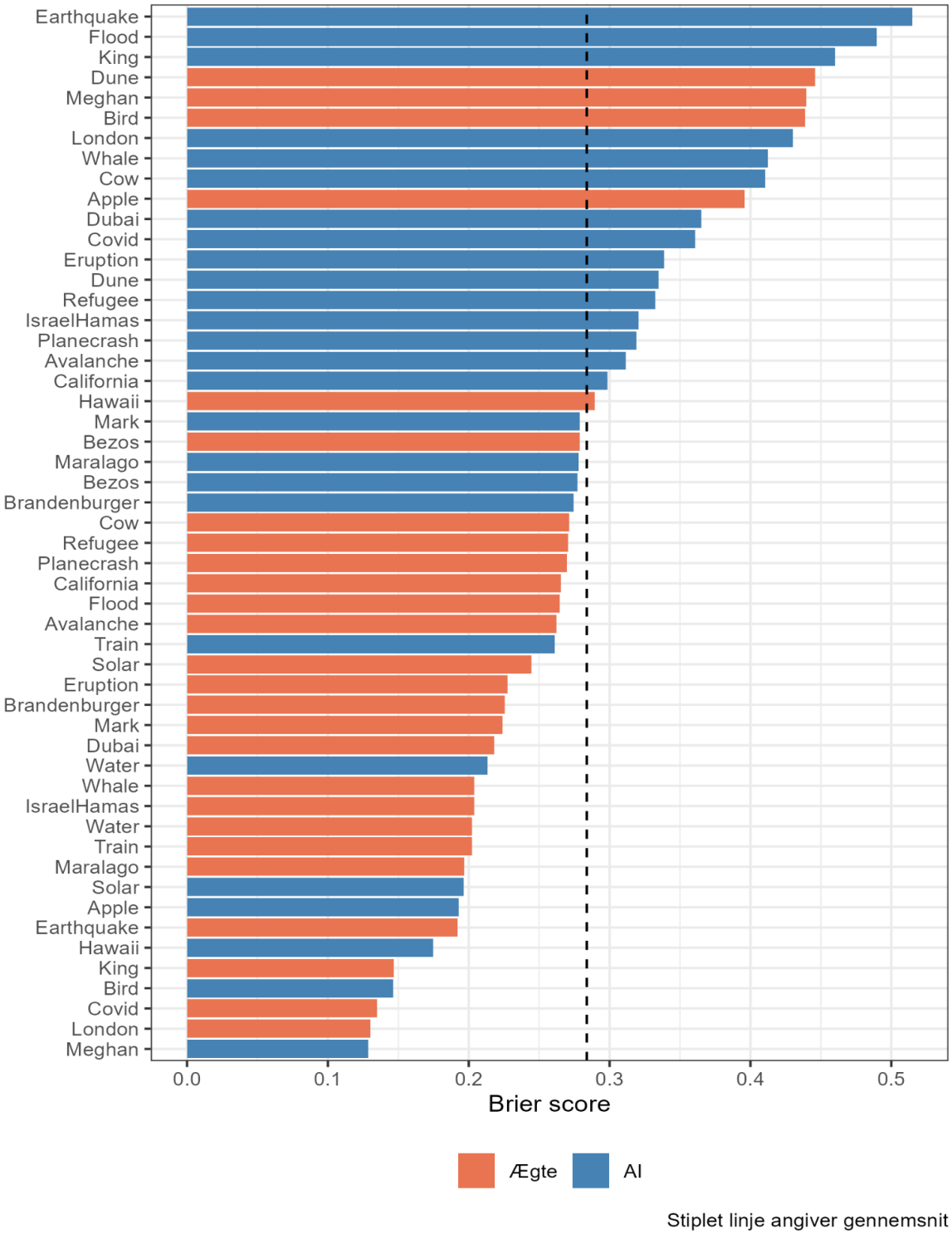
I og med at respondenterne tilfældigt er blevet tildelt forskellige billeder kan vi ligeledes tjekke, hvor stor forskel der er på tværs af billeder i forhold til, hvor lette eller svære de har været at vurdere ægtheden af. Ligeledes kan det fortælle os noget om, hvorvidt respondenterne har haft sværest ved at vurdere ægtheden af ægte eller AI-skabte billeder.

Figur 3C2 viser gennemsnitlige brier scores pr. billede. Som det ses er der stor forskel på, hvor svært respondenterne har haft ved at vurdere billederne, hvor bl.a. et AI-skabt billede af Meghan Markle har været let at vurdere, mens et AI-skabt billede af en oversvømmelse har været svært at vurdere ægtheden af.

Det ses ligeledes, hvordan det generelt er AI-skabte billeder, der har høje brier scores, hvilket indikerer, at det er primært AI-skabt indhold, respondenterne har været svært ved at vurdere ægtheden af, mens de bedre har styr på at vurdere ægte billeder.

Det er værd at bemærke, at på trods af at respondenterne har modtaget forskellige billeder, der varierer i sværhedsgrad, har dette hverken været muligt at vurdere a priori ligesom de ikke burde give anledning til nogen forskydninger i fordelingerne, da det snarere efterligner virkeligheden, hvor nogle AI-billeder er godt og andre minde gode. Derudover introducerer variationen ikke *bias* i vores regressioner. Det skyldes at forskelle i sværhedsgrad er tilfældige målefejl på den afhængige variabel, hvilket ikke introducerer *bias,* men blot inefficiens.

Figur 3C2: Brier scores pr. billede

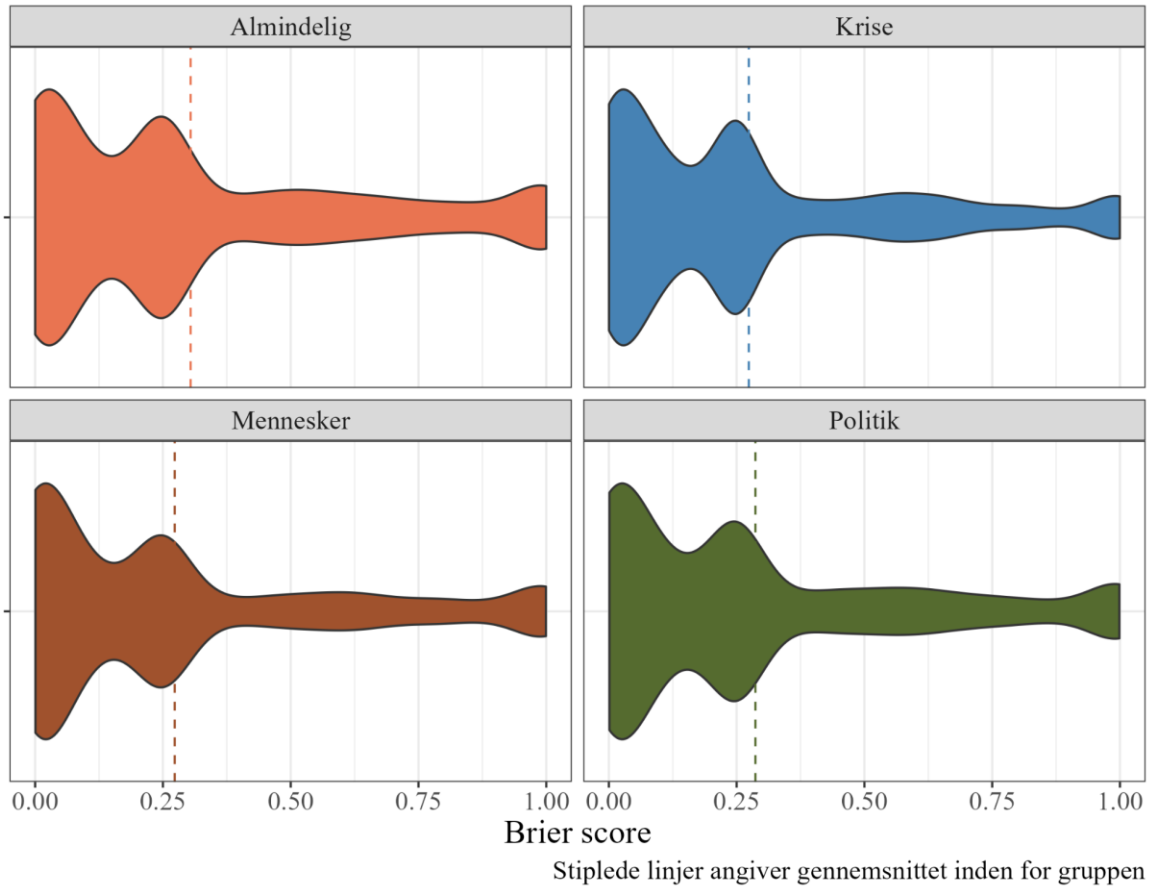


Respondenternes evne til at idenficere forskellige typer af billeder afhængigt af dets tema (krise, politisk, mennesker og almindelig) er ligeledes visualiseres. Dette ses i figur 3C3, som viser fordelingen af brier scores pr. kategori af billeder.

Figuren viser tydeligt, hvordan danskerne har lige svært ved at identificere, hvorvidt et billede er ægte, uanset billedkategori. Det ses ved, at de fire fordelinger i høj grad ligner hinaden og at gennemsnittet, markeret med stiplede linjer, ligger tæt på hinanden. Dette kan synes overraskende da f.eks. billeder, der viser mennesker også er svære for respondenterne at aflæse, selvom vi omgås mennesker hver eneste dag og derfor forventeligt har en ret god fornemmelse for menneskelige detaljer.

Figuren har således potentielt en anden vigtig politisk implikation, i den forstand at den indikerer at det ikke er let at lave reguleringsmæssige rækværk mod specifikke typer af billeder, f.eks. regulering af hvad AI-genereret indhold må afbillede af hensyn til, hvor svært det er for brugerne at aflæse. Det synes lige så svært at identificere et AI-billede af et flystyrt som af en ko, der græsser.

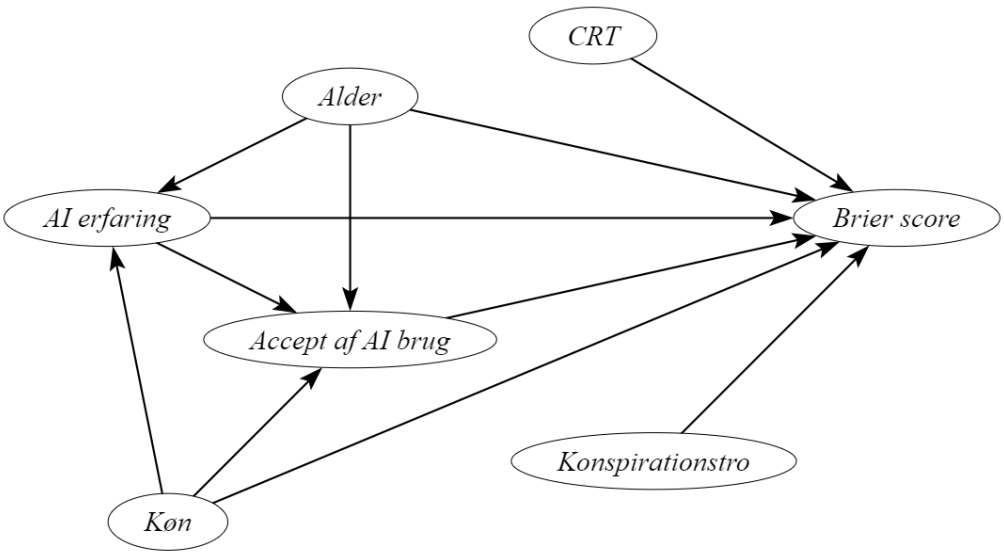
Figur 3C3: Fordeling af brier scores pr. kategori af billeder



Bilag 3D: Specificering af regressionsmodeller af AI-billede eksperiment

Forskellene mellem stimuligrupper er udregnes med OLS regression. Modellerne er konfigureret efter kausaldiagrammet neden for.

Det betyder, at to af modellerne inkluderer statistisk kontrol. Det gælder sammenhængen mellem henholdsvis erfaring med AI og dygtigheden til at identificere ægtheden af billeder, og sammenhængen med accept af AI som uafhængig variabel. I førstnævnte model kontrolleres der for respondentens alder og køn. I sidstnævnte model kontrolleres der for alder, køn og erfaring med AI.



Bilag 3E: Design af misinformationsopmærksomhed eksperiment

Eksperimentet er designet til at skulle identificere effekten af forskellige udlægninger af de samfundsmæssige udfordringer med misinformation. Specifikt hvorvidt graden af alarmisme påvirker respondenternes frygt for at blive eksponeret for misinformation, frygt for at misinformation påvirker demokratiet og demokratiske valg negativt, frygt for at blive beskyldt for at sprede misinformation og opbakning til demokratiet som styreform.

Eksperimentet har tre stimuli grupper. Følgende stimuli er tildelt de forskellige grupper:

Affejende information: *Det er svært at sige, hvor stor en udfordring misinformation er. Misinformation har i nogle lande vist sig at være et mindre problem, hvor forskning fra USA viser, hvordan mængden af misinformation ofte er mindre end frygtet. Vi ved imidlertid ikke meget om, hvorvidt det samme gør sig gældende i Danmark eller om den situation ændrer sig, nu hvor generativ kunstig intelligens kan bruges til at generere misinformation.*

Balanceret information: *Det er svært at sige, hvor stor en udfordring misinformation er. Misinformation har i nogle lande vist sig at være et mindre problem, hvor forskning fra USA viser, hvordan mængden af misinformation ofte er mindre end frygtet. Vi ved imidlertid ikke meget om, hvorvidt det samme gør sig gældende i Danmark eller om den situation ændrer sig, nu hvor generativ kunstig intelligens kan bruges til at generere misinformation.valg i 2024.*

Alarmistisk information: *Frygten for spredning af misinformation har taget til de senere år, særligt i forbindelse med udbredelsen af generativ kunstig intelligens ("AI"), der hurtigt og virkelighedsnært kan generere misinformation. Derfor har World Economic Forum også udpeget netop misinformation som den største globale trussel på kort sigt, særligt i lyset af at knap halvdelen af verdens befolkning skal stemme til demokratiske valg i 2024.*

Dernæst er alle respondenter bedt om at svare på følgende spørgsmål (skala 1-5 fra meget enig til meget uenig):

- Information jeg ser på sociale medier er generelt til at stole på*
- Traditionelle nyhedsmedier deler ikke misinformation på sociale medier
- Folkevalgte politikerne deler ikke misinformation når de skriver på sociale medier
- Generativ kunstig intelligens ("AI") er positivt for det danske demokrati*
- Misinformation påvirker, hvem der vinder valg til f.eks. Folketinget*
- Demokratier er ubeslutsomme og har for meget politisk fnidder*
- At have eksperter i stedet for politikere til at tage beslutninger for, hvad de synes er bedst for landet vil være en god måde at lede det her land på*
- Partier som repræsenterer folk som mig bliver ofte fejlagtigt beskyldt for at sprede misinformation
- Jeg frygter i politiske diskussioner med venner og familie at blive beskyldt for at sprede misinformation

Spørgsmål 1-3 indgår i et indeks om respondentens frygt for at blive eksponeret for misinformation. Spørgsmål 4-5 indgår i et indeks om respondentens frygt for negative demokratiske effekter af misinformation og AI. Spørgsmål 6-7 indgår i et indeks om respondentens opbakning til demokrati som styreform. Spørgsmål 8-9 indgår i et indeks om respondentens frygt for at blive beskyldt for at sprede misinformation.

* angiver hvorvidt forskellen mellem den alarmistiske og den affejende gruppe er statistisk signifikant på et 95% konfidensinterval.