ET KIG IND I DEN SORTE BOKS

En analyse af sociale mediers begrænsning af skadeligt og ulovligt indhold i 2025

/Digitalt Ansvar

INDLEDNING

På sociale medier kan brugere dele næsten hvad som helst, når som helst på døgnet. For at beskytte mod skadeligt og ulovligt indhold fjerner eller begrænser de sociale medier i mere eller mindre grad indhold fra deres brugere. Det kaldes også moderation. Uden den form for moderation ville platformene flyde over med spam, porno, svindel og skadeligt indhold. Det ville hurtigt få både brugere og annoncører til at forlade sociale medier. Moderation er derfor definerende for vores oplevelse og sikkerhed på digitale platforme.

Alligevel viser undersøgelser igen og igen, at brugere eksponeres for skadeligt og ulovligt indhold på sociale medier. Ifølge Medierådet og Center for Sociale Medier, Tech og Demokrati har to ud af fem unge i alderen 13-17 år inden for det seneste år set eller oplevet indhold online, som de fandt ubehageligt[1]. Samtidig har Meta i starten af 2025 svækket sine platformspolitikker og sin håndhævelse[2].

Moderation på digitale platforme er derfor stadig en politisk kampplads, hvor vinde fra USA skubber i retning af mindre moderation, mens EU forsøger at trække udviklingen i den modsatte retning. Med Digital Services Act (DSA) har vi i EU vedtaget lovgivning, som skal beskytte borgerne mod ulovligt og skadeligt indhold online. Lovgivningen forpligter desuden de store sociale medier til at offentliggøre, hvad de fjerner, hvor meget de fjerner, og hvordan de fjerner indhold på platformene. Disse tal offentliggøres bl.a. i EU-Kommissionens gennemsigtighedsdatabase, som Digitalt Ansvar i denne analyse er dykket ned i.

Hvis vi ønsker at beskytte både børn og voksne mod skadeligt og ulovligt indhold, er moderation af indhold stedet at starte. Vi må forstå, hvordan moderationen fungerer i praksis, og hvordan de forskellige sociale medier løfter deres ansvar.

Formålet med analysen er derfor at sammenligne TikTok, Instagram, Facebook og Snapchat på en række parametre, der kaster lys over deres moderation i praksis.

OPSUMMERING

Analysen bygger på data fra EU-Kommissionens gennemsigtighedsdatabase og sammenligner, hvordan fire store sociale medier, TikTok, Facebook, Instagram og Snapchat, modererer indhold. Formålet er både at belyse forskelle i platformenes praksis og at vise, hvad databasen kan og ikke kan bruges til.

Analysen dokumenterer betydelige forskelle i platformenes moderationspraksis, både hvad angår mængde, metoder og hastighed.

Analysen peger på fem centrale konklusioner:

1. **TikTok begrænser og fjerner over 89 gange så meget indhold som Snapchat:** Analysen viser markante forskelle i, hvor meget indhold de forskellige platforme fjerner og begrænser. TikTok skiller sig markant ud ved at moderere mest indhold og Snapchat mindst.

2. **TikTok modererer mest proaktivt, og Snapchat mindst:** Platformene bruger også meget forskellige metoder til at moderere indhold. TikTok anvender i høj grad automatiserede systemer til at opdage og moderere indhold, mens Snapchat i langt højere grad er afhængig af anmeldelser fra brugerne. Denne forskel tyder på at have en betydelig effekt på, hvor meget indhold platformene modererer.

3. **Snapchat er langsomst til at fjerne eller begrænse indhold:** Der er betydelige forskelle i den gennemsnitlige tid, det tager platformene at fjerne og begrænse det ulovlige og skadelige indhold, de opdager. Snapchat er den platform, der er langsomst til at fjerne indhold, mens TikTok er hurtigst.

4. **Nye tendenser inden for indholdsmoderation:** Sammenligner man begyndelsen af 2025 med årets slutning, står det klart, at Instagram både modererer mindre indhold automatisk og i mindre grad fjerner og begrænser indhold. Denne udvikling tyder på, at de ændringer i indholdsmoderationen, som Meta annoncerede i januar 2025, er trådt i kraft.

5. **Sociale medier slører omfanget af ulovligt indhold:** Når de fire platforme fjerner indhold, begrunder de det i langt de fleste tilfælde med, at indholdet strider mod platformenes egne regler, i stedet for at indholdet er ulovligt ifølge national lovgivning. Dette reducerer gennemsigtigheden og vanskeliggør vurderingen af, hvor meget ulovligt indhold der rent faktisk findes og fjernes.

Hvad er EU-Kommissionens gennemsigtighedsdatabase?

FAKTASIDE

1.

Gennemsigtighedsdatabasen er et offentligt register, der indeholder begrundelser for beslutninger om at moderere indhold fra digitale tjenester ("statements of reasons"). Det er digitale tjenester med status som mellemstore eller store online platforme eller meget store online platforme (VLOP's), som er forpligtede til at indlevere disse til databasen.

2.

Når en digital tjeneste modererer indhold, skal den sende en begrundelse til brugeren (DSA, artikel 17). Disse begrundelser skal også sendes til Kommissionen uden unødig forsinkelse, så de kan indgå i databasen (DSA, artikel 24, stk. 5). Formålet med databasen er at "sikre gennemsigtighed og muliggøre kontrol med beslutninger om indholdsmoderation truffet af udbydere af onlineplatforme og overvågning af udbredelsen af ulovligt indhold på internettet" (DSA, betragtning 66)[3].

3.

Databasen suppleres af andre indsigtsmuligheder, såsom forskeradgang for "godkendte forskere" (DSA, artikel 40, stk. 4) og de årlige gennemsigtighedsrapporter, der bl.a. indeholder antal modtagne klager og overordnede oplysninger om tjenesternes indholdsmoderation (DSA, artikel 15). Begrundelserne for fjernelse af indhold, som indsendes til databasen, supplerer således den årlige gennemsigtighedsrapport samt de yderligere oplysninger, som godkendte forskere i teorien kan få adgang til.

Hvordan modereres indhold på sociale medier?

Arbejdet med at holde sociale medier fri for ulovligt og skadeligt indhold bliver udført af både maskiner, mennesker ansat af platformene selv, brugere og pålidelige indberettere ("trusted flaggers"). Alle disse aktører indgår i den proces, der kaldes indholdsmoderation.

Hvad er indholdsmoderation?

Indholdsmoderation er den proces, hvor brugergenereret indhold overvåges, filtreres, rangeres og eventuelt fjernes på sociale medier for at sikre, at platformenes regler og gældende lovgivning overholdes.

Menneskelige moderatorer er ansat af de sociale medier til at behandle anmeldelser fra brugere og indhold, som er blevet "flagged" af platformens egne AI-systemer. De udfører et nødvendigt og skjult arbejde ved at gennemse ulovligt og skadeligt indhold. For år tilbage spillede de en større rolle end i dag, men med COVID-nedlukningen blev mange moderatorer sendt hjem, og mere automatiserede løsninger tog over. Den automatiserede indholdsmoderation blev samtidig set som en løsning på den stigende mængde indhold, der skal modereres.[4]

Fordelen ved menneskelig moderation er, at mennesker bedre forstår konteksten, fx hvis noget er ment som humor eller satire. Derfor kan de bedre vurdere, hvorvidt indhold, der tilsyneladende strider mod platformens retningslinjer, måske alligevel ikke skal fjernes. Mennesker er ligeledes bedre til hurtigt at tilpasse sig ændringer i platformenes retningslinjer end automatiske systemer.

Ulempen ved menneskelig moderation er, at der uploades og anmeldes så meget indhold på de sociale medier hvert sekund, at det er umuligt for moderatorerne at følge med. Dertil er det potentielt skadeligt for disse mennesker at blive eksponeret for traumatiserende indhold hver dag, såsom halshugningsvideoer og overgrebsmateriale.

Automatisk indholdsmoderation foregår ved at anvende teknologi såsom kunstig intelligens, nøgleord og hashing til at identificere skadeligt og ulovligt indhold og beslutte, om indholdet skal fjernes.

Nogle systemer kan finde og fjerne nye versioner af indhold, der allerede er blevet fjernet før, men som uploades igen, eller indhold, der i en anden kontekst er blevet erklæret ulovligt. Et eksempel på dette er *StopNCII*, der anvender perceptual hashing, en form for "digitalt fingeraftryk", der automatisk kan identificere, hvis registreret intimt billedmateriale deles uden samtykke på en af de platforme, der er tilmeldt *StopNCII*.

Andre systemer er trænet til at identificere nyt materiale. Her vil der fx blive anvendt superviseret maskinlæring. Træning af et sådant system foregår ved, at en algoritme vises en lang række eksempler, fx på homofobiske udsagn, og på den baggrund lærer, hvad der kendetegner denne type

udsagn. På den måde kan systemet selv klassificere, hvorvidt nyt indhold indeholder homofobiske udsagn. Et sådant system vil ikke være 100 procent korrekt, så der vil altid være et menneskeligt valg i, hvor sikkert systemet skal være, før indholdet kommer til kontrol hos et menneske eller automatisk bliver fjernet. Hvis systemet indstilles til at fjerne indhold, som vurderes at være ulovligt med 90 procent sikkerhed, vil meget af det ulovlige indhold blive fjernet, men noget lovligt indhold vil også blive fjernet. Hvis en platform derimod vælger kun at fjerne indhold, som systemet er 99,9 procent sikker på er ulovligt, vil meget lidt indhold blive fjernet ved en fejl, men meget ulovligt indhold vil også slippe igennem og fejlagtigt blive klassificeret som lovligt. Dette forhold mellem *falske negative* og *falske positive* er i sidste ende derfor ikke kun en teknisk udfordring, men også et ideologisk spørgsmål, der beror på en afvejning af vores fundamentale rettigheder, fx ytrings- og informationsfrihed over for retten til privatliv eller beskyttelse mod hadtale.

Fordelen ved automatisk moderation er både den hastighed og den skala, hvormed ulovligt og skadeligt indhold kan blive fjernet på sociale medier. Disse systemer arbejder langt hurtigere end menneskelige moderatorer, og de kan potentielt fjerne indhold, før personer eksponeres for det.

Ulempen ved disse automatiske systemer er, at risikoen for at implementere en systematisk fejlagtig indholdsmoderation stiger sammenlignet med anvendelsen af menneskelige moderatorer, selvom mennesker formodentlig heller ikke modererer fejlfrit. Ligeledes kan AI være mindre fleksibel end menneskelige moderatorer, hvis der fx sker ændringer i platformenes indholdspolitikker, på trods af at disse teknologiske løsninger konstant forbedres.[5]

Brugere kan anmelde indhold til sociale medier, som de mener overskrider platformens retningslinjer eller national og europæisk lovgivning. I DSA'en beskriver artikel 16, hvordan denne "notice-and-action"-mekanisme skal stilles til rådighed for alle brugere af lagringstjenester, herunder online platforme.

Pålidelige indberettere er organisationer, som af nationale DSA-tilsyn er blevet givet denne status. Statussen indebærer en form for forrang, hvor anmeldelser fra pålidelige indberettere prioriteres af de sociale medier, fordi det formodes, at anmeldelserne har en højere kvalitet end normale brugeres. Det vil typisk være eksperter fra civilsamfundet, der får denne status. Ifølge DSA (artikel 22) er det *ikke* påkrævet, at digitale tjenester *handler* på anmeldelser fra pålidelige indberettere, kun at anmeldelserne prioriteres og vurderes hurtigt. I Danmark er Red Barnet og Rettighedsalliancen pålidelige indberettere i henhold til DSA.

Metode

Denne analyse bygger på tal fra EU-Kommissionens gennemsigtighedsdatabase indrapporteret i 2025. Visse af analysens tal dækker hele perioden, hvor data er aflæst fra databasens dashboard. Andre dele af analysen baserer sig på data fra seks udvalgte uger i 2025, da datapunkterne ikke fremgår af databasens dashboard. Derfor er data i stedet blevet downloadet, behandlet og analyseret i Python. De seks udvalgte uger er uge 4, 14, 24, 36, 43 og 50. Disse data er blevet downloadet fra gennemsigtighedsdatabasens hjemmeside den 8. april 2026. Det vil fremgå tydeligt af brødteksten i de afsnit, hvor analyserne kun er baseret på disse udvalgte uger.

Der vil kun indgå TikTok-data for andet halvår af 2025, da platformen modtog kritik i en uafhængig audit, om at de indrapporterede i dubletter til databasen indtil maj 2025, hvor de efter at være blevet gjort opmærksom på fejlen, har ændret deres procedurer [6]. For at gøre TikToks tal sammenlignelige med de øvrige platformes helårstal omregnes tallene for andet halvår til estimerede helårstal. Dette gøres ved at gange antallet af moderationsindgreb i perioden 1. juli–31. december med 365/184 svarende til cirka 1,98. Beregningen forudsætter, at antallet af moderationsindgreb i første halvår ville have ligget på omtrent samme niveau som i andet halvår. Der tyder på stadig at være en usikkerhed forbundet med de indrapporterede tal fra Tiktok, da der er store udsving og manglende data for visse dage i andet halvår af 2025. Det er ikke muligt at vurdere udefra, hvad disse udsving skyldes. På grund af Tiktok's varierende indberetninger til databasen vil data fra uge 35, 42 og 51 anvendes som de udvalgte uger. Forskellene mellem de forskellige moderationspraksisser på tværs af de sociale medier er så store, at vi har valgt at medtage TikTok i analysen på trods af denne usikkerhed.

Moderation af profiler, grupper, hashtags etc. indgår ikke i denne analyse, som kun omfatter moderation af billeder, video, tekst og lyd. Desuden er datapunkter, som er blevet modereret under en såkaldt "monetær restriktion", sorteret fra, da dette som udgangspunkt ikke påvirker brugernes oplevelse af de sociale medier, men blot hvorvidt indholdsproducenter kan tjene penge på et stykke indhold.

Data fra de sociale medier TikTok, Instagram, Facebook og Snapchat er blevet analyseret. Indhold modereres på flere forskellige måder: Et moderationsindgreb kan være, at indholdet slettes eller bliver aldersbegrænset, men det kan fx også blive algoritmisk nedprioriteret. At indhold nedprioriteres algoritmisk betyder, at det bliver mindre synligt i fx feeds og søgeresultater, uden at det dog bliver fjernet helt.

Det er ikke muligt kun at fokusere på indhold publiceret fra Danmark, så analysen baserer sig på data fra hele EU. Da digitale tjenester kan indsende begrundelser til databasen med en forsinkelse, kan der være tilføjet nye begrundelser til databasen, siden data er downloadet. Disse ekstra begrundelser er af et relativt begrænset omfang og påvirker derfor ikke analysens overordnede pointer.

ANALYSE

1. TikTok fjerner mest, og Snapchat fjerner mindst

Tabel 1 nedenfor viser forskellige nøgletal om det samlede antal moderationsindgreb på tværs af de forskellige platforme. Nøgletallene præsenteres som absolutte tal og i procent, hvor det er relevant.

Tabel 1. Overblik over indholdsmoderation

	TikTok	Instagram	Snapchat	Facebook
Antal stykker indhold, der er blevet modereret	332.576.923	21.556.333	3.737.663	92.066.263
Andel indhold modereret som er blevet slettet	46,2% (153.640.880)	71,7% (15.452.167)	77,8% (2.907.559)	72,0% (66.285.988)
Modereret med artikel 16 som kilde: "notice and action"	0,04% (117.843)	0,75% (160.867)	68,3% (2.554.473)	0,37% (339.011)
Modereret med pålidelige indberettere som kilde	0,0001% (323)	0,0009% (191)	0,008% (311)	0,0008% (724)
Frivillig moderation	99,9% (332.457.346)	99,3% (21.395.275)	31,6% (1.181.487)	99,6% (91.726.528)

Tabel 1 viser, hvordan TikTok modererer betydeligt mere indhold end andre sociale medier. De har fx modereret 89 gange mere indhold på deres platform end Snapchat og 15,5 gange mere end Instagram.

Den altoverskyggende kilde til moderationsindgreb er platformenes egen frivillige moderation, da kun få stykker indhold er blevet modereret på grund af artikel 16 i DSA'en og ordningen om pålidelige indberettere (DSA, artikel 22).

At TikTok fjerner så meget mere indhold end de andre platforme, kan formodentlig til dels forklares med, at der uploades mere offentligt indhold* på denne platform. Der er dog ingen officielle tal om dette.

* DSA'en skelner mellem onlineplatforme, der lagrer og udbreder brugerindhold til offentligheden, og interpersonelle kommunikationstjenester. Forskelle i, hvor stor en del af brugen der består af offentligt tilgængeligt indhold frem for privat kommunikation, kan derfor påvirke sammenligningen mellem platformene.

Det kan også skyldes, at der simpelthen er mere indhold at moderere, hvorfor sandsynligheden for mere skadeligt og ulovligt indhold også stiger. Det er dog usandsynligt, at dette forhold forklarer hele forskellen, for målt på antal aktive månedlige brugere i EU er både Instagram og Facebook større end Tiktok, og TikTok er ikke engang dobbelt så stor som Snapchat.

Så disse forskelle mellem sociale medier indikerer en forskel i moderationspolitikker og -praksisser, men et referencetal, som det samlede antal uploadede stykker indhold pr. dag, ville øge transparensen yderligere, da man derved kunne opgøre, hvor stor en andel af det samlede uploadede indhold der modereres.

"Platformen har fx modereret 89 gange mere indhold end Snapchat og 15,5 gange mere end Instagram "

Tabel 2. Månedlige brugere i EU ifølge VLOPs egne gennemsigtighedsrapporter.

	TikTok	Instagram	Snapchat	Facebook
Antal månedlige brugere i EU	169.000.000[7]	281.800.000[8]	94.810.865[9]	263.600.000[10]

2. Pålidelige indberettere er kilde til få moderationsindgreb

Tabel 1 viser, at ordningen med pålidelige indberettere kun fører til meget få moderationsindgreb. Det skyldes formentlig, at der i øjeblikket eksisterer to parallelle ordninger: en ældre frivillig model og den nye DSA-baserede ordning, som i Danmark administreres af DSA-tilsynet. Indtil videre har kun Rettighedsalliancen og Red Barnet opnået officiel status som pålidelige indberettere.

Derfor kan ordningens fulde potentiale endnu ikke vurderes ud fra gennemsigtighedsdatabasen. For at øge transparensen og styrke ordningen opfordres flere organisationer til at søge status gennem DSA-tilsynet. Dette kan støttes politisk, fx ved at gøre det økonomisk attraktivt at deltage i ordningen.

Samtidig viser tallene, at selv på EU-niveau er det kun en forsvindende lille andel af fjernet indhold, der stammer fra pålidelige indberettere. Ordningen løser altså ikke det strukturelle problem med skadeligt og ulovligt indhold på sociale medier, men fungerer primært som et vigtigt værktøj til at sikre hurtig handling i enkeltsager.

3. "Notice and action"-mekanismen mangler stadig at komme op i gear

Artikel 16 i DSA'en er den såkaldte "notice and action"-mekanisme. Den pålægger digitale tjenester at have lettilgængelige og brugervenlige anmeldelsesmekanismer, så brugere kan indberette ulovligt indhold på platformen. Platformene er pålagt at behandle alle anmeldelser, som de modtager gennem disse mekanismer, og træffe deres beslutning "rettidigt, omhyggeligt, uden forskelsbehandling og objektivt". Dette giver brugere flere rettigheder over for digitale tjenester, men placerer samtidig også et medansvar hos dem, da de skal bidrage til at holde digitale tjenester ansvarlige. Brugere er dermed en vigtig aktør, hvis reguleringen skal blive en succes.

Som det fremgår af Tabel 1, er der få stykker ulovligt indhold, som modereres på baggrund af en anmeldelse hjemlet i artikel 16. Det tyder på, at brugere mangler viden om deres nye rettigheder, samt hvordan de i praksis anvendes.

Undtagelsen er Snapchat, hvor 68,3 procent af deres modererede indhold har artikel 16-anmeldelser som kilde. Dette er muligvis også et udtryk for, at Snapchat fjerner mindre indhold proaktivt end de andre platforme, og at platformen registrerer alle modtagne anmeldelser som artikel 16-anmeldelser.

4. Sociale medier modererer på baggrund af egne servicevilkår

Som vist i Tabel 3 nedenfor registrerer tjenesterne som udgangspunkt størstedelen af det modererede indhold som værende i strid med deres servicevilkår (artikel 17, 3e), når de fjerner det, i stedet for at kategorisere det som ulovligt indhold (artikel 17, 3d). Dette er også tilfældet for Snapchat, selvom platformen bliver gjort opmærksom på en stor del af det modererede indhold via artikel 16-anmeldelser, som ellers netop drejer sig om ulovligt indhold.

Denne praksis gør det svært at bruge databasen til at få indblik i, hvor meget ulovligt indhold der faktisk findes på platformene – selvom det netop er et af databasens formål (DSA, betragtning 66).

Hvad denne praksis skyldes, er ikke til at sige, men to muligheder fremstår plausible: 1) De digitale tjenester vil gerne have, at det fremstår, som om der er mindre ulovligt indhold på deres platforme, end tilfældet er. 2) De digitale tjenester vurderer, at det er en større byrde at registrere indholdet som lovovertrædelser, da de også er pålagt at registrere, hvilke love de mener, det specifikke stykke indhold har overtrådt. Det er lettere for dem at behandle og registrere overtrædelser af deres egne servicevilkår, fordi de samme regler gælder i hele EU[11] – hvilket også gør det nemmere at automatisere processen.

Tabel 3. Andel af indhold, som er modereret, fordi det er registreret som ulovligt.

	TikTok	Instagram	Snapchat	Facebook
Indhold som er modereret, fordi det er ulovligt	0,0007% (88165)	0,013% (2718)	0,010% (392)	0,01% (10.468)

5. TikTok modererer mest proaktivt, og Snapchat mindst

Indholdsmoderation på sociale medier består af både automatisk og menneskelig moderation. Det gælder både *identifikationen* af potentielt skadeligt og/eller ulovligt indhold, og selve *beslutningen* om at fjerne det. Indholdsmoderationen foregår derfor i to mere eller mindre automatiserede skridt.

Hvor meget de forskellige sociale medier anvender automatiseret indholdsmoderation, varierer kraftigt fra platform til platform.

I Tabel 4 kan vi se, hvor stor en andel af det samlede indhold der henholdsvis detekteres automatisk og manuelt.

Tabel 4. Automatisk identifikation af indhold

	Andel automatisk identificeret	Andel menneskelig identificeret
Snapchat	18,8% (701.032)	81,2% (3.036.631)
Facebook	82,5% (75.998.883)	17,5% (16.067.380)
Instagram	91,4% (19.707.900)	8,6% (1.848.433)
TikTok	97,8% (325.348.894)	2,2% (7.228.030)

Af tabellen fremgår det tydeligt, at TikTok identificerer mest indhold automatisk i både absolutte og relative tal. Snapchat er dog det sociale medie, der skiller sig mest ud, da platformen i langt mindre grad detekterer indhold automatisk. Dette faktum, sammenholdt med at Snapchat i absolutte tal fjerner klart mindst indhold, indikerer, at *Snapchat i betydeligt mindre grad proaktivt modererer på deres platform end andre sociale medier.*

Tabel 5. Automatisk moderation af indhold

	Andel automatisk eller delvist automatisk begrænset	Andel menneskelig begrænset
Snapchat	18,0% (672.884)	82,0% (3.064.779)
Instagram	91,4% (19.707.900)	8,6% (1.848.433)
Facebook	82,5% (75.988.883)	17,5% (16.067.380)
TikTok	92,9% (309.090.742)	7,1% (23.486.181)

Et lignende billede tegner sig, når vi undersøger, hvordan de sociale medier beslutter, hvorvidt indhold skal modereres. Her anvender TikTok og Instagram i høj grad automatiske metoder, mens Snapchat hovedsageligt benytter sig af menneskelig moderation.

Indhold identificeret automatisk

SNAPCHAT

18,8%

TIKTOK

97,8%

6. Snapchat er langsomst om at moderere indhold

Dette afsnit baserer sig ikke på data fra hele 2025, men på seks udvalgte uger for Snapchat, Instagram og Facebook og tre uger for TikTok (se metodeafsnittet for yderligere oplysninger).

Moderation af indhold kan foregå hurtigt efter, at indholdet er blevet uploadet, hvis platformenes kunstige intelligenssystemer opdager overtrædelser af platformens servicevilkår eller lovgivningen. Men indholdet kan også ligge på platformen i lang tid, inden det fjernes. Tiden mellem, at indhold uploades til platformen, og at det fjernes, er blevet kaldt *moderationsforsinkelsen*[12]. Dette nøgletal er vigtigt, da flere mennesker potentielt kan udsættes for skadeligt og ulovligt indhold, jo længere tid indholdet ligger på platformen.

Tabel 6 opgør forskellige nøgletal for moderationsforsinkelsen på tværs af platformene målt i antal dage. For at sikre et mere lige sammenligningsgrundlag mellem platformene – Facebook har fx været tilgængelig i Danmark i en del længere tid end TikTok – er indhold, der er uploadet for mere end tre år siden, frasorteret analysen.*

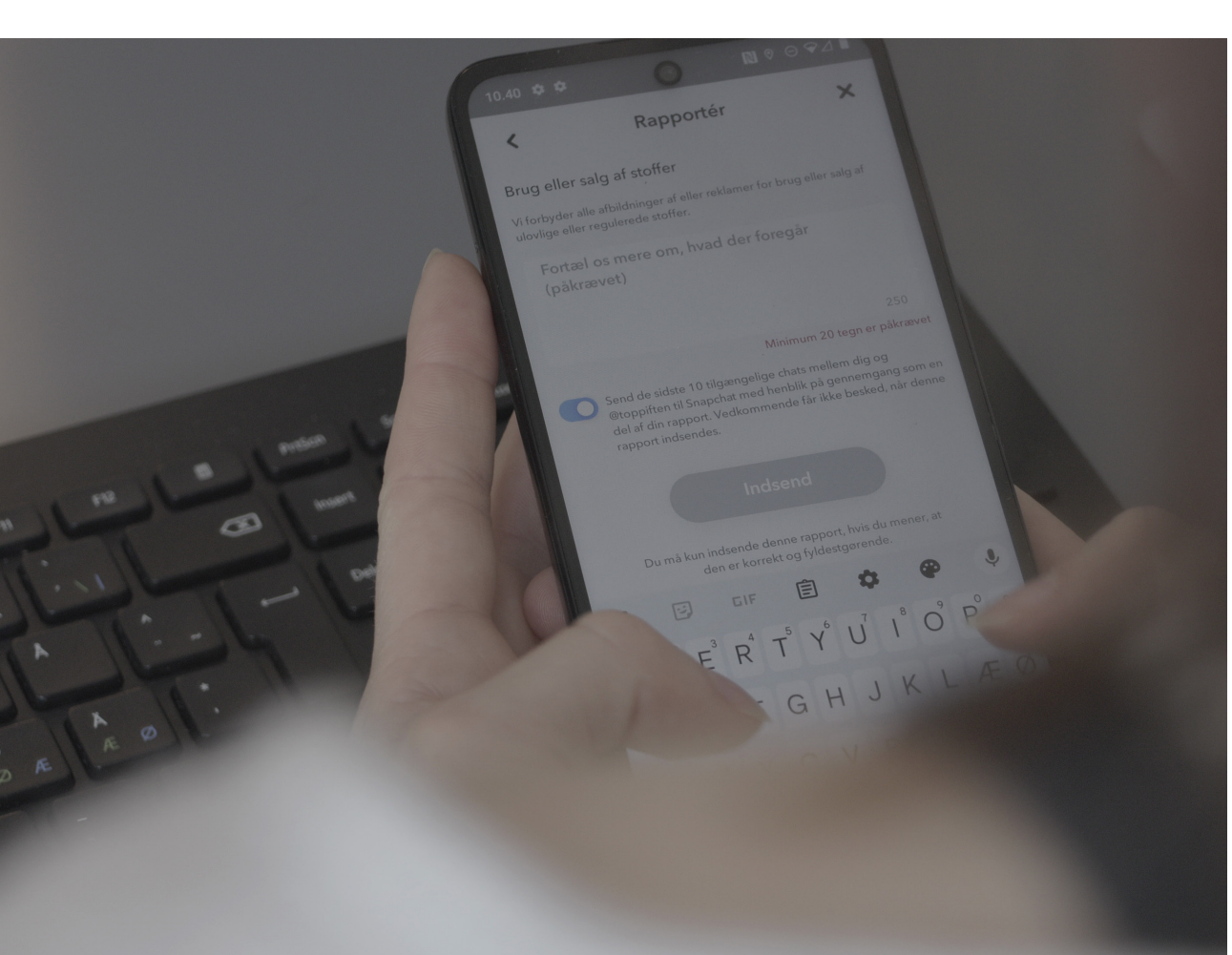
Tabel 6. Moderationsforsinkelsen

	Median moderationsforsinkelse	Gennemsnitlig moderationsforsinkelse	Standardafvigelsen for moderationsforsinkelsen
TikTok	0 dages forsinkelse	12 dages forsinkelse	67,5 dage
Instagram	0 dages forsinkelse	13 dages forsinkelse	82 dage
Facebook	0 dages forsinkelse	16,5 dages forsinkelse	90 dage
Snapchat	0 dages forsinkelse	18 dages forsinkelse	95,5 dage

* Alt indhold, der ligger mere end 1095 dage før den første dag i hver af de seks analyseperioder, er frasorteret. Moderationsforsinkelsen er udregnet på baggrund af alle seks perioder.

Tabellen viser, at der er stor forskel på moderationsforsinkelsen på tværs af platformene. TikTok er hurtigst til at fjerne indhold, mens Snapchat er langsomst målt på den gennemsnitlige moderationsforsinkelse.

Median-moderationsforsinkelsen på alle platforme er 0 dage, hvilket indikerer, at en stor del af indholdet modereres samme dag, som det uploades. Forskellene i den gennemsnitlige moderationsforsinkelse og standardafvigelsen skyldes sager med en stor moderationsforsinkelse. Det samme billede tegner sig, når vi visualiserer fordelingen af moderationsforsinkelser for hvert af de sociale medier på en log-log-skala, hvor alle værdier på 0 er inkluderet som 0,5 for at kunne blive vist på denne skala (Bilag A). Som det fremgår af visualiseringerne, har moderationsforsinkelsen for alle de sociale medier en lang hale. Denne lange hale viser, at selvom meget indhold bliver modereret hurtigt, er der stadig en ikke-ubetydelig del, som tager meget lang tid at få fjernet. Dette gør sig gældende for alle fire sociale medier.



7. Hvis indhold ikke automatisk identificeres første dag, er især Snapchat afhængig af brugeranmeldelser

Dette afsnit baserer sig ikke på data fra hele 2025, men på seks udvalgte uger for Snapchat, Instagram og Facebook og tre uger for TikTok (se metodeafsnittet for yderligere oplysninger).

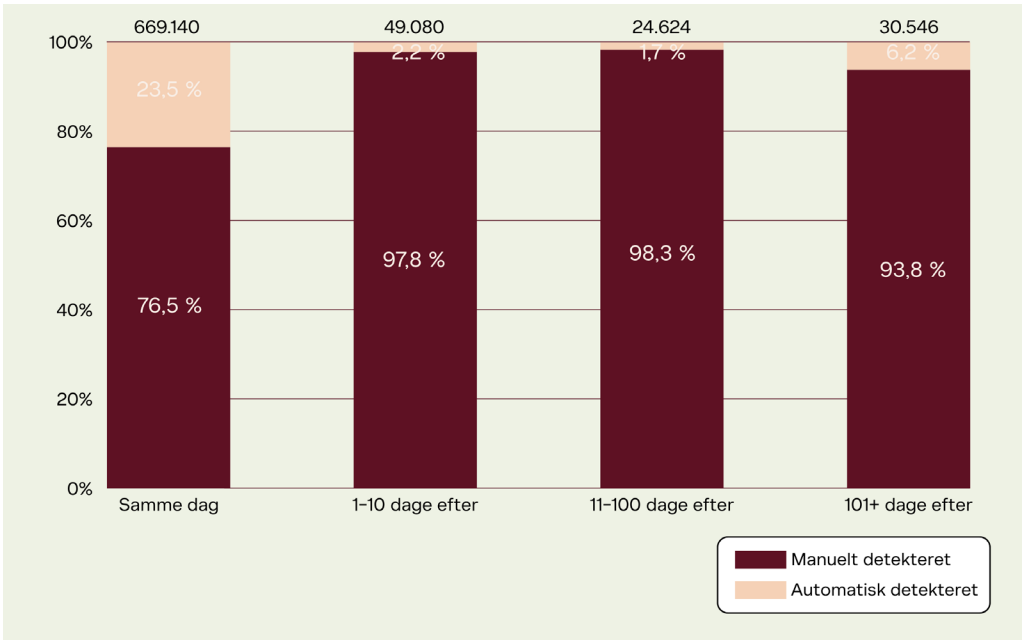
Nedenstående figurer viser andelen af modereret indhold, der er detekteret henholdsvis automatisk og manuelt, opdelt efter hvor lang tid der går fra upload til moderation. Tallene over søjlerne angiver antallet af fjernelser i hvert tidsinterval.

På tværs af figurerne er det tydeligt, at automatiske metoder til detektion af skadeligt eller ulovligt indhold spiller en stor rolle samme dag, som det uploades. Det tyder på, at platformenes automatiske systemer, med undtagelse af Snapchat, er i stand til at identificere og fjerne en stor andel af problematisk materiale, men at denne effektivitet falder markant, hvis indholdet ikke bliver fanget med det samme.

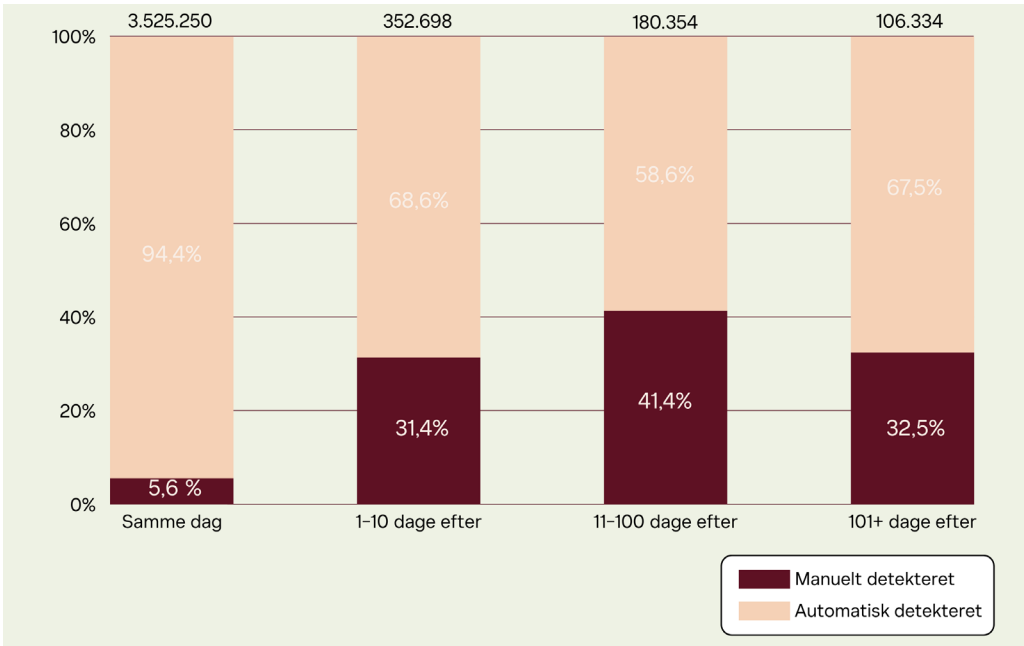
Når indholdet ikke bliver identificeret automatisk den første dag, er der markante forskelle mellem platformene i, hvor meget de efterfølgende afhænger af manuel moderation. Snapchat skiller sig særligt ud ved i høj grad at være afhængig af brugeranmeldelser og manuel detektion. Instagram fjerner også mere indhold efter brugeranmeldelser sammenlignet med Facebook og TikTok, hvor den automatiske indholdsmoderation er forholdsvis konstant, uanset hvor længe indholdet har ligget på platformen.

”Når indholdet ikke bliver identificeret automatisk den første dag, er der markante forskelle mellem platformene i, hvor meget de efterfølgende afhænger af manuel moderation.”

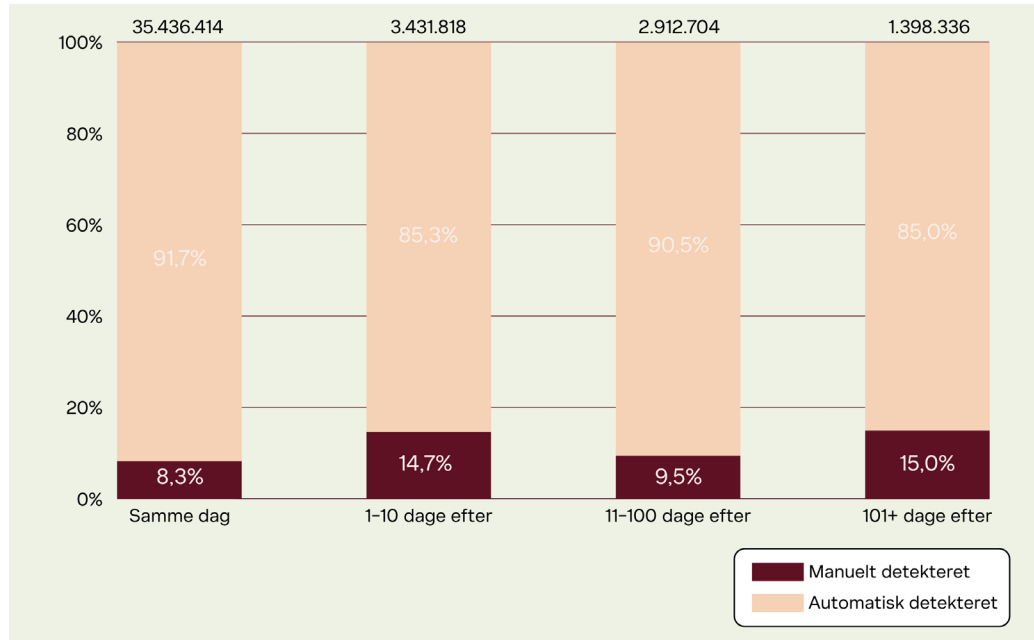
Figur 1. Fjernelse af indhold efter detektionstype – Snapchat



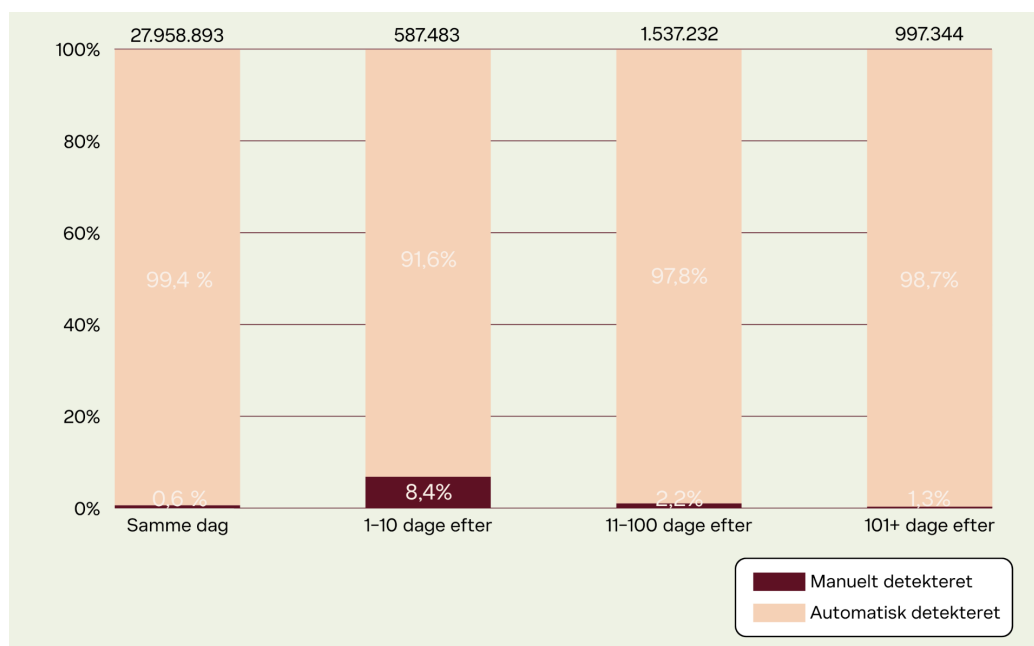
Figur 2. Fjernelse af indhold efter detektionstype – Instagram



Figur 3. Fjernelse af indhold efter detektionstype – Facebook



Figur 4. Fjernelse af indhold efter detektionstype – TikTok



8. Instagram har ændret moderationspraksis

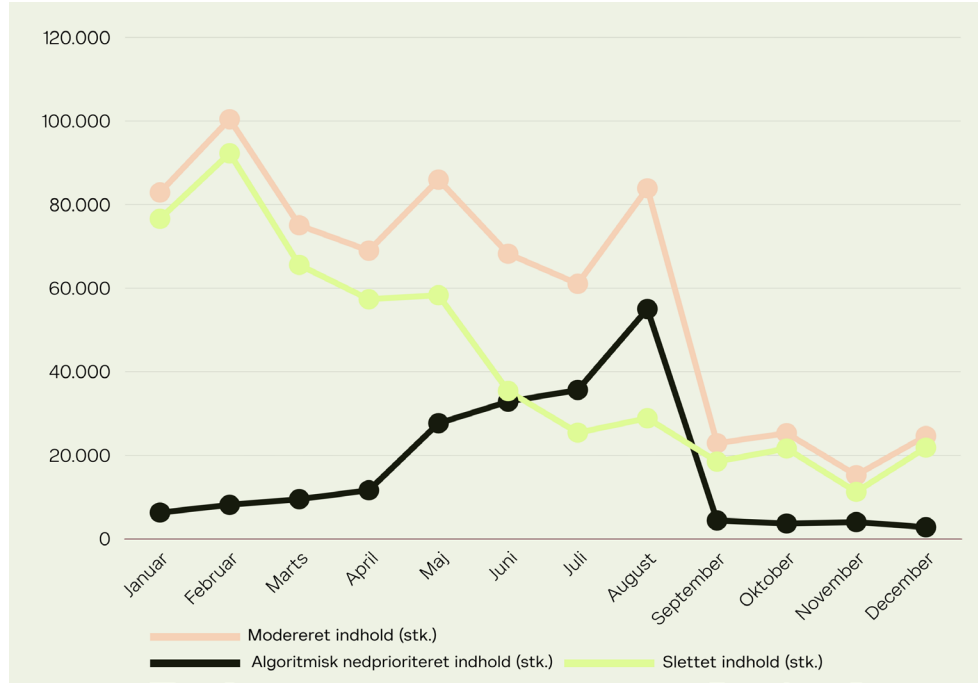
"Selvom udviklingen gennem året er volatil, ses der alligevel en markant moderationsnedgang i årets sidste måneder."

Undersøger man platformenes gennemsnitlige daglige moderation gennem 2025, ser man, at Instagram lader til at have ændret moderationspraksis.

Som Figur 5* på næste side viser, begrænsede og fjernede Instagram i gennemsnit 82.904 stykker indhold dagligt i januar 2025, mens platformen i december kun modererede 24.626 stykker indhold. Selvom udviklingen gennem året er volatil, ses der alligevel en markant moderationsnedgang i årets sidste måneder. Sammenligner man første kvartal med fjerde kvartal 2025, er der tale om et fald på *hele 74 procent.*

Frem til juni er særligt udviklingen i forholdet mellem slettet og algoritmisk nedprioriteret indhold bemærkelsesværdig. Som det fremgår af Figur 5, bliver en markant mindre andel af det modererede indhold fjernet over tid, mens vi ser en stigende andel algoritmisk nedprioritering. Fra august falder dog også mængden af indhold, der algoritmisk nedprioriteres.

Figur 5. Gennemsnitlig daglig moderation – Instagram



Ændringerne kan formodentlig forklares med Metas nye strategi, som blev præsenteret i januar 2025[13]. Her meddelte virksomheden, at den fremover vil gøre op med den overmoderation, den hævder at have praktiseret de seneste år. Meta mener at have modereret en for stor mængde indhold, som virksomhedens automatiske systemer har identificeret, og at virksomheden fremover vil ændre praksis.

Policy-ændringerne indebærer blandt andet færre restriktioner på politisk indhold, samtidig med at deres automatiserede systemer fremadrettet skal fokusere mere på at håndtere ulovligt indhold og alvorlige overtrædelser, herunder terrorisme, narkotikahandel, svindel og seksuel udnyttelse af børn.

Det er nærliggende at tolke udviklingen i Instagrams moderationspraksis som et tegn på, at Metas annoncerede ændringer nu er trådt i kraft. Vi ser derfor de første tegn på, at den kulturkamp om indholdsmoderation, som den amerikanske regering og tech-giganter har indledt, også har en praktisk effekt for europæiske brugere.

* For Periode 2 er 11 stykker af det modererede indhold kategoriseret som "begrænset". Dette udgør kun 0.0026% af det totale indhold og er derfor ikke inkluderet i Figur 5.

DISKUSSION OG PERSPEKTIVER

Hvad vi kan og ikke kan sige ud fra modereret indhold

Denne analyse bygger på data fra EU-Kommissionens gennemsigtighedsdatabase. Det er dog vigtigt at være opmærksom på, at disse data kun beskriver det indhold, som platformene selv har vurderet som værende i strid med lovgivningen og/eller deres egne servicevilkår, og som derfor er blevet modereret. Det betyder, at tallene potentielt underrapporterer mængden af skadeligt og ulovligt indhold af to grunde:

_01 De sociale medier modererer ikke indhold, som de burde moderere

En undersøgelse foretaget af tre forskere fra IT-Universitetet fra 2024 finder for eksempel, at kun 247 ud af 1130 anmeldelser af hadtale på sociale medier blev fjernet[14]. Digitalt Ansvar har også vist, at Snapchat ikke fjerner åbenlyse narkoprofiler, selvom platformen modtager en anmeldelse[15].

Da gennemsigtighedsdatabasen kun indeholder information om indhold, der er blevet modereret, er den bedre i stand til at analysere potentiel overmoderation end undermoderation: Det skyldes, at vi ikke kan se, hvor meget indhold der er blevet detekteret eller anmeldt, men som i sidste ende ikke er blevet fjernet.

_02 De sociale medier kan kun fjerne det indhold, de kender til

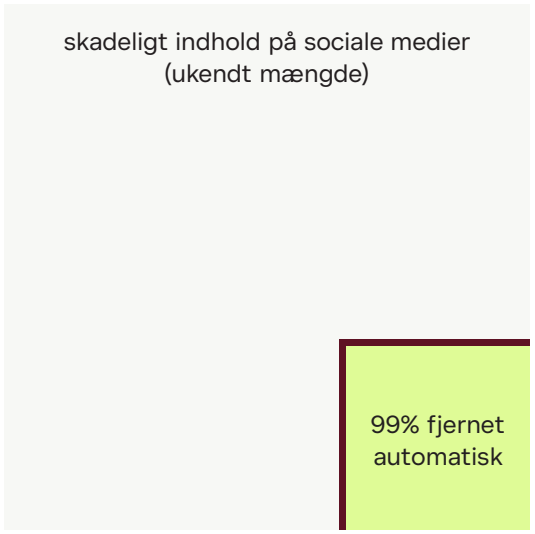
Når sociale medier får kritik for manglende indholdsmoderation, forsvarer de sig ofte med at fremlægge deres proaktiv-rate. Det vil sige den mængde indhold, som de har detekteret og fjernet, inden de har modtaget en brugeranmeldelse om selvsamme stykke indhold. Her ligger djævlen dog i detaljen, for proaktiv-raten er udregnet ud fra det indhold, som de kender til og har fjernet.

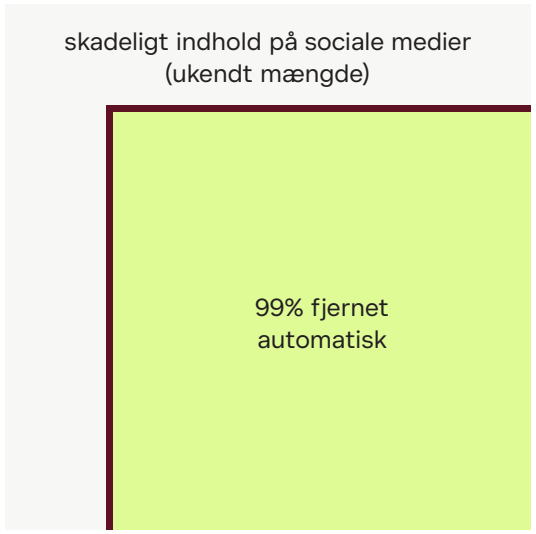
Det samme gør sig gældende for denne analyses resultater. Når Instagram fx identificerer og modererer X procent selvskademateriale på sin platform, før det bliver indrapporteret, er det altså ikke det samme, som at de fjerner X procent af alt selvskademateriale på sin platform proaktivt – *for Instagram kender ikke til alt selvskademateriale på sin platform*. Dette viste Digitalt Ansvar med al tydelighed i analysen InstaHarm[16].

Vigtigheden af denne nuance kan illustreres med nedenstående figur.

- Den gule firkant repræsenterer den totale mængde skadeligt og/eller ulovligt indhold på et givent socialt medie.
- Den røde firkant repræsenterer det indhold, som det sociale medie har fjernet.
- Den grønne firkant viser den del, som det sociale medie selv har detekteret og fjernet, dvs. proaktiv-raten.

I begge figurer har det sociale medie fjernet 99 procent af det skadelige og/eller ulovlige indhold på sin platform proaktivt, men i figuren til højre gør platformen et relativt godt stykke arbejde, mens den i figuren til venstre gør et dårligt stykke arbejde.





Så når denne analyses resultater fx viser, at TikTok fjerner mest indhold proaktivt, er det ikke *nødvendigvis* et udtryk for en mere sikker platform, da den samlede mængde skadeligt og/eller ulovligt indhold på platformen er ukendt. Og det er derfor også direkte forkert, når TikToks nordiske chef udtaler til Berlingske, at "mere end 99 procent af det indhold, der ikke skal være på appen, bliver fjernet, før det overhovedet bliver vist"[17]. For hun kender slet ikke til alt det indhold, der ikke skal være på appen.

Konklusion

Denne overordnede analyse af EU-Kommissionens gennemsigtighedsdatabase har givet flere forskellige indsigter:

Mængden af modereret indhold varierer enormt: Analysen viser, at sociale medier håndterer indholdsmoderation på vidt forskellige måder. TikTok skiller sig markant ud ved at moderere meget mere indhold end de andre platforme i absolutte tal. Dette kan enten skyldes strengere politikker eller en større mængde uploadet skadeligt og ulovligt indhold. Denne forskel kan også være en indikator på, at TikTok har bedre eller mere aggressive automatiserede systemer til indholdsmoderation. Snapchat ligger i den modsatte ende af skalaen, da Snapchat er den online platform, der modererer mindst.

Automatiseret versus manuel moderation: TikTok, Facebook og Instagram anvender automatiseret indholdsmoderation i meget høj grad sammenlignet med Snapchat, der lader til at basere sin moderation langt mere på brugeranmeldelser. Denne forskel tyder på at have en betydelig effekt på, hvor meget indhold platformene modererer, da Snapchat også fjerner betydeligt mindre indhold i absolutte tal. Det er dog værd at have in mente, at Snapchat adskiller sig fra de andre sociale medier ved, at private beskeder fylder mere end offentlige opslag, hvilket påvirker platformens mulighed for at anvende automatiseret moderation.

Tidsforskelle i moderationsforsinkelse: Data viser, at der er betydelige forskelle i den gennemsnitlige tid, det tager platformene at fjerne indhold. TikTok er hurtigst til at moderere indhold, mens Snapchat har den længste gennemsnitlige moderationsforsinkelse.

Median-moderationsforsinkelsen for alle platforme er dog 0 dage, hvilket indikerer, at en stor del af indholdet modereres hurtigt på tværs af alle platforme. Dog har alle de sociale medier også en ikke-ubetydelig andel indhold, som har ligget på deres platform i lang tid, inden det er blevet opdaget og modereret. Der er derfor stadig behov for løbende forbedring af systemer til indholdsmoderation.

Ændret moderationspraksis i 2025: Hvis man ser på den årlige udvikling for Instagram, lader det til, at platformen har ændret moderationspraksis. Det afspejles i årets første halvdel ved, at en betydelig mængde indhold bliver algoritmisk nedprioriteret i stedet for at blive fjernet. Fra august reducerede platformen også den mængde indhold, der blev algoritmisk nedprioriteret, og platformen modererede overordnet set markant mindre end i starten af året. Ændringerne kan tyde på, at Metas nye moderations-strategi, som blev præsenteret i januar 2025, er trådt i kraft.

Databasens begrænsninger

Mangel på data om undermoderation: Databasen fokuserer udelukkende på det indhold, der fjernes eller modereres, men den giver ikke indsigt i, hvor meget skadeligt og ulovligt indhold der aldrig opdages og fjernes, eller som anmeldes, men ikke modereres.* Dette er en væsentlig blind vinkel, der kan påvirke vores forståelse af platformenes sikkerhedsniveau. Sat på spidsen er det ud fra databasens informationer ikke til at svare på, om TikTok er mere eller mindre sikker for børn og unge, fordi platformen fjerner langt mere indhold: Er det et udtryk for en stor mængde skadeligt og ulovligt indhold på TikTok, som skal fjernes, eller er det et udtryk for, at platformen er bedre til at rydde op?

Fravær af referencepunkter: For at kunne sætte mængden af fjernet indhold i perspektiv mangler vi data om, hvor meget indhold der uploades i alt på platformene, og hvor mange anmeldelser der indgives, uanset om de er hjemlet i artikel 16 eller ej. Dette ville gøre det muligt at beregne forholdet mellem det modererede indhold og den samlede mængde indhold på platformene samt andelen af anmeldelser, der fører til et moderationsindgreb.[18]

* I platformenes gennemsigtighedsrapporter fremgår det dog, hvor mange artikel 16-anmeldelser de har fået, og hvor mange af dem der leder til en begrænsning. Dog tyder resultaterne på, at fortolkningen af artikel 16-anmeldelser varierer fra platform til platform, hvor Meta og TikTok i højere grad differentierer mellem eksplicitte artikel 16-anmeldelser og almindelige brugeranmeldelser efter platformens egne retningslinjer. Det vanskeliggør en direkte sammenligning, hvorfor en sådan ikke er foretaget.

NOTER

1 Medierådet for børn og unge. (2025). Undersøgelse: Unges brug af digitale medier. https://medieraadet.dk/viden/mediebrug-og-trivsel/undersoegelse-unges-brug-af-digitale-medier#:~:text=Dette%20vises%20i%20en%20ny%20unders%C3%B8gelse%20af%20unges,bruger%20digitale%20medier.%20Besvarelserne%20er%20indsamlet%20af%20Voxmeter.

2 Kaplan, J. (2025, 7. januar). More Speech and Fewer Mistakes. Meta. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

3 Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act). https://www.eu-digital-services-act.com/Digital_Services_Act_Preamble_61_to_70.html

4 Gillespie, T. (2020). Content moderation, AI, and the question of scale. Big Data & Society. https://doi.org/10.1177/2053951720943234

5 Willats, R., Pennington, J., Mohan, A. & Vidgen, B. (2025). Classification is a RAG problem: A case study on hate speech detection. arXiv. https://arxiv.org/abs/2508.06204

6 KPMG (2025). DSA Assurance Report: Independent practitioner's assurance report concerning Regulation (EU) 2022/2065, the Digital Services Act (DSA), for TikTok Technology Limited. DSAAssuranceReport-TikTokTechnologyLimited-2025.pdf

7 TikTok. (2025). European Union (EU) – Monthly Active Recipients Report. https://www.tiktok.com/transparency/en/eu-mau-2025-1.

8 Meta. (2025, 29. august). Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Instagram.

9 Snapchat. (2025). Transparency. https://values.snap.com/privacy/transparency/european-union

10 Meta. (2025, 29. august). Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Facebook. https://transparency.meta.com/reports/regulatory-transparency-reports/

11 Keller, D. (2025, 2. september). A Primer on Cross-Border Speech Regulation and the EU's Digital Services Act. Center for Internet and Society at Stanford Law School. https://cyberlaw.stanford.edu/blog/2025/09/a-primer-on-cross-border-speech-regulation-and-the-eus-digital-services-act/

12 Kaushal, R., Van De Kerkhof, J., Goanta, C., Spanakis, G., & Iamnitchi, A. (2024). Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 1121-1132).

13 Kaplan, J. (2025, 7. januar). More Speech and Fewer Mistakes. Meta. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

14 Kristensen, M.V. (2024). Undersøgelse: Sociale medier fjerner ikke ulovlig hadtale. IT-Universitetet i København. https://itu.dk/Om-ITU/Presse/Nyheder/2024/Sociale-medier-fjerner-ikke-ulovlig-hadtale?.

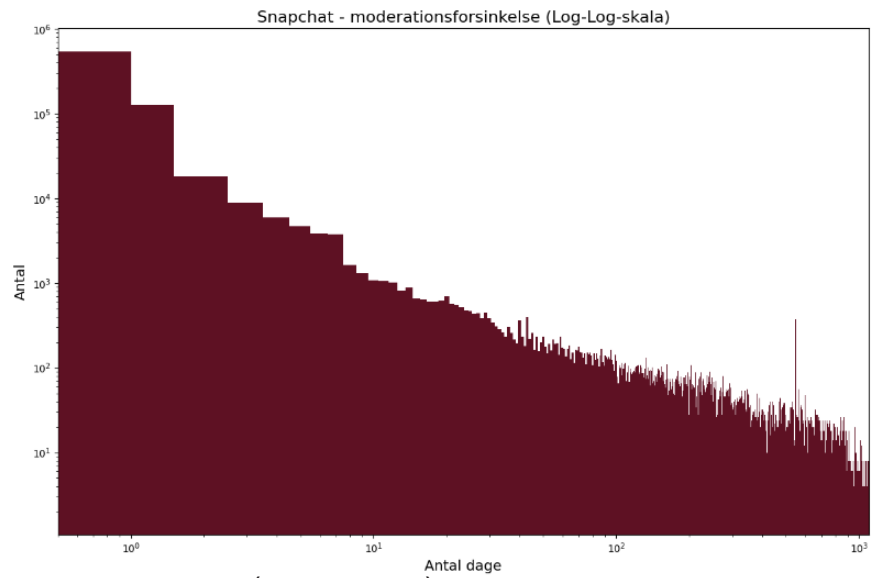
15 Digitalt Ansvar. (2025). Ny undersøgelse: Snapchat ignorerer anmeldelser af danske narkoprofiler og anbefaler dem endda til børn. https://www.digitaltansvar.dk/aktuelt/snapchat-ignorerer-anmeldelser-af-danske-narkoprofiler

16 Digitalt Ansvar. (2024). Instaharm. En undersøgelse af Instagrams manglende indholdsmoderation af selvskadeindhold. https://drive.google.com/file/d/1MZrFRii_nJYdW8RulORB9JveLkCRbncX/view?pli=1

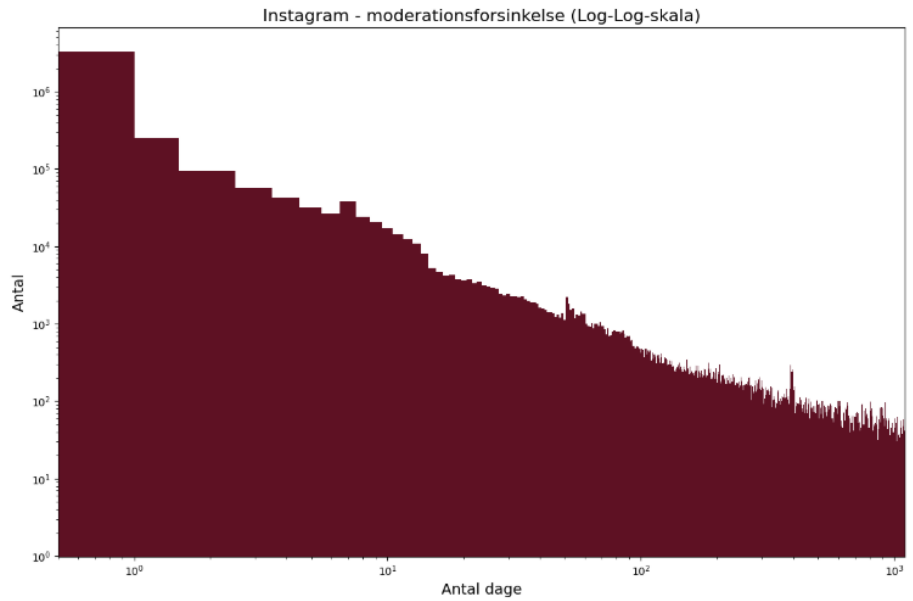
17 Sandberg, D-M., Larsen, J.L., Kjems-Krognos, N. (2024, 14. november). Kritikken er haglet ned over TikTok – men nordisk chef affejer anklagerne: ›Min TikTok-oplevelse er 100 procent positiv‹. Berlingske. https://www.berlingske.dk/virksomheder/kritikken-er-haglet-ned-over-tiktok-men-nordisk-chef-affejer?gaa_at=eafs&gaa_n=AerBZYO03Qtn8_CBRzAaWrb7M_UErXc5V3wRk5p9zdL7KwU2YhBLg-rlRX4O7tD8u0Q%3D&gaa_ts=67d2fc7f&gaa_sig=X7H6ieZisTnBrYbdyc4951S5I679KxKMzi-FGC9qalUwEgcQx4G31KL4m4G2HYARtwsR723YclqkNNWDk1z23Q%3D%3D

18 Groesch, S., Birrer, A., Just, N. & Saurwein, F. (2026). Big data, small answers: How the DSA Transparency Database falls short of its regulatory objectives. Telecommunications Policy, Vol. 50. No. 1. https://doi.org/10.1016/j.telpol.2025.103088

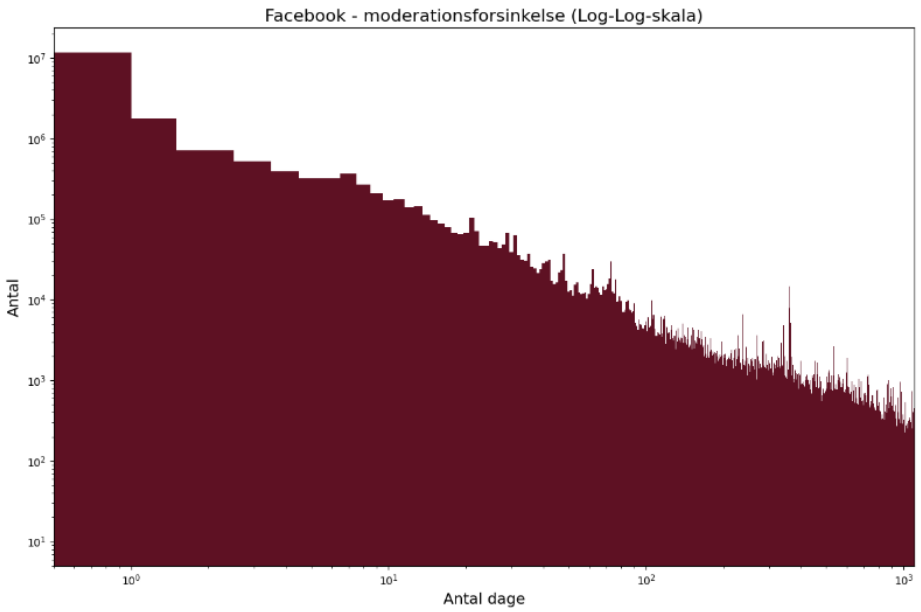
Bilag A. Moderationsforsinkelse (Log-Log Skala) efter platform



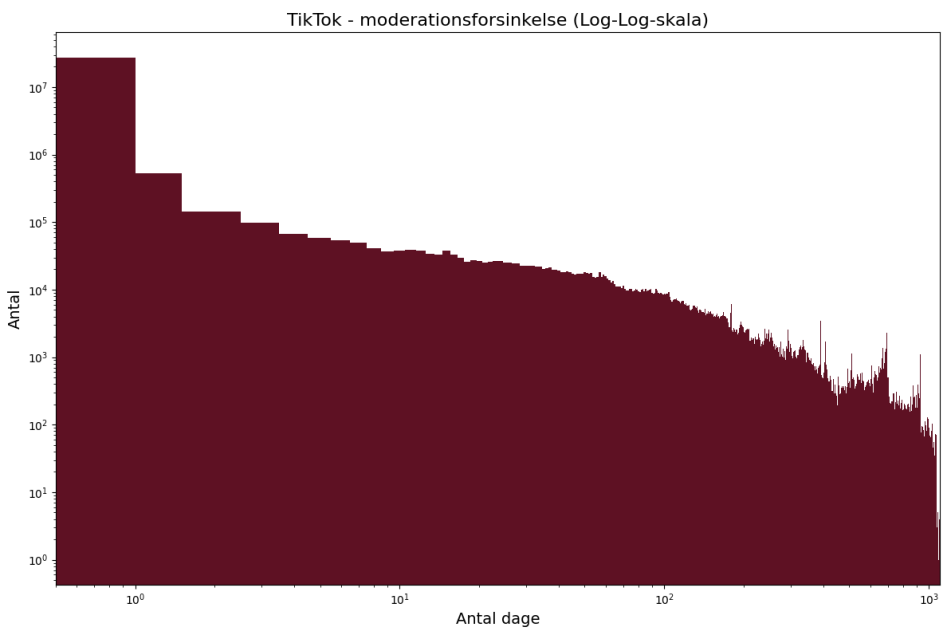
Figur A1. Moderationsforsinkelse (Log-Log Skala) – Snapchat.



Figur A2. Moderationsforsinkelse (Log-Log Skala) – Instagram.



Figur A3. Moderationsforsinkelse (Log-Log Skala) – Facebook.



Figur A4. Moderationsforsinkelse (Log-Log Skala) – TikTok.

/Digitalt Ansvar

Udgivet af Digitalt Ansvar, juni 2026
Ansvarshavende redaktør: Ask Hesby Holm
Redigering: Asger Nim og Asta Iris Rohde
Tekst og analyse: Åsa Vigen Bardal og Asger Nim

/Digitalt Ansvar