/Digitalt Ansvar

Analyse

InstaHARM

En undersøgelse af Instagrams manglende indholdsmoderation af selvskadeindhold.

Indholdsfortegnelse

Indledning

Sociale medier har i mange år fået kritik for ikke at gøre nok for at forhindre spredning af selvskadeindhold. Det kan være indhold, der viser, romantiserer eller instruerer til selvskade. I mange tilfælde er også anbefalingsalgoritmerne med til at sprede indhold og profiler om selvskade.[1] Allerede i 2017 fik Instagram kritik, efter den 14-årige Molly Russell tog sit eget liv. Molly havde i seks måneder op til sin død set, gemt og liket cirka 12 Instagramopslag om dagen relateret til selvmord, selvskade og depression. [2]

Også herhjemme har Instagram mødt kritik. DR-dokumentaren "Døde pigers dagbog" viste i 2020, hvordan Instagram faciliterede kontakt mellem selvskadende unge, som delte farlige råd og opfordringer til selvskade, og hvor enkelte begik selvmord. [3]

Sociale medier som Instagram har forsvaret sig med, at de hele tiden investerer i ny og bedre teknologi og er blevet bedre til at fjerne skadeligt indhold.[4] I 2020 offentliggjorde Instagram, at man ville anvende teknologi til at fjerne åbenlyse eksempler på selvskade.[5]

I 2024, i kølvandet på DR's dokumentar "Alene hjemme på internettet", udtalte politisk chef for Meta, selskabet bag Facebook og Instagram, at de "er blevet meget bedre til det [at fjerne skadeligt indhold], end vi var for tre til fem år siden." [6]

Metas fælleskabsregler beskriver desuden, at man fjerner alt indhold, som opfordrer til selvmord, selvskade eller spiseforstyrrelser, uanset materialets kontekst. [7]

Undersøgelsen viser, at Instagram ikke fjerner selvskadeindhold

I denne analyse har vi undersøgt, om Instagram er blevet bedre til at fjerne selvskadeindhold, og hvorvidt der er indbygget sikkerhed ift. deres anbefalingsalgoritme, så den ikke udbreder kendskabet til selvskadegrupper.

For at undersøge dette oprettede vi et netværk af private Instagramprofiler. Fra profilerne uploadede vi 85 opslag, der indeholdt selvskaderelateret materiale, der gradvis blev mere og mere åbenlyst selvskadende. Opslagene udtrykte ønske om selvskade, delte råd om selvskadende adfærd, viste billeder af stadig mere alvorlig selvskade og opfordrede andre medlemmer af netværket til selvskade.

- Instagrams moderation af selvskadeindhold er særdeles mangelfuld. Instagram fjernede ikke et eneste opslag, der blev delt. Selv ikke åbenlyst selvskaderelateret indhold, der tydeligt viser blod, barberblade eller opfordring til selvskade, blev fjernet.

- Digitalt Ansvar testede, om kunstig intelligens var i stand til automatisk at identificere selvskadematerialet, vi uploadede på Instagram. Algoritmen, vi benyttede, identificerede 38 procent af selvskadebillederne. Ud af de billeder, vi uploadede på Instagram, der mest åbenlyst viste selvskade, identificerede algoritmen hele 88%. Dette peger på, at Instagram har *valgt* ikke at benytte teknologi, som kan minimere problemet med selvskademateriale på platformen.

- Instagrams anbefalingssystem anbefalede andre 13-årige profiler at blive venner med de resterende profiler i selvskadegruppen, hvis de 13-årige profiler blev venner med bare én bruger i selvskadegruppen. Anbefalingsalgoritmen kan dermed være medvirkende til at forme og udbrede netværk, hvor der deles selvskademateriale.

Denne analyse indeholder åbenlys omtale af selvskade. Har du brug for hjælp, eller kender du nogen, der har, kan du kontakte:

Livslinien på tlf. 70 201 201 mellem kl. 11 og 05 alle ugens dage.
Læs mere på Livslinien.dk

Foreningen Spiseforstyrrelser og Selvskade på tlf. 7010 1818.
Deres telefonrådgivning har åbent hver mandag og torsdag
kl. 9.00-19.00 samt tirsdag og onsdag kl. 16.00-19.00.
Læs mere på Spiseforstyrrelse.dk

Undersøgelsen peger på, at Meta ikke lever op til EU-lovgivningen

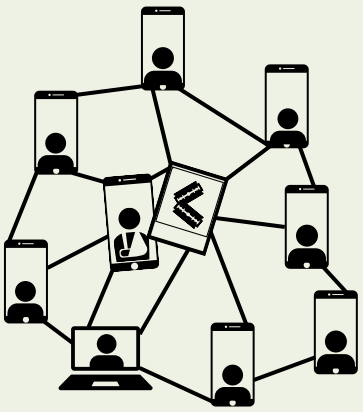
Den særdeles mangelfulde moderation af selvskadeindhold og profiler tyder på, at platformen ikke handler i overensstemmelse med Forordningen om Digitale Tjenester (The Digital Services Act, herfra DSA). Ifølge DSA'en er de store digitale tjenester forpligtet til at identificere og begrænse systemiske risici, der blandt andet omfatter forventede negative konsekvenser for mental og fysisk helbred samt børns beskyttelse.

_01 Instagram overholder ikke egne anvendelsesvilkår om indholdsmoderation (DSA, artikel 14, 4), hvor det står beskrevet, at billeder med fysisk selvskade fjernes proaktivt ved hjælp af kunstig intelligens.

_02 Instagrams manglende indholdsmoderation er et udtryk for, at platformen ikke i tilfredsstillende grad beskytter brugerne mod systemiske risici. Digitalisering af selvskadeindhold udgør en risiko for brugernes fysiske og mentale helbred samt mindreåriges sikkerhed, som Instagrams indholdsmoderation ikke formår at begrænse (DSA, artikel 34-35).

_03 Instagrams anbefalingssystem, der anbefaler andre Instagrambrugere, udgør i sig selv en systemisk risiko mod mindreåriges sikkerhed og brugernes fysiske og mentale helbred, da systemet hjælper brugere med at identificere selvskadenetværk. Meta har en forpligtelse til at begrænse denne risiko ved for eksempel at ændre, hvordan deres anbefalingsalgoritme fungerer, sådan den ikke forstærker disse selvskadenetværk (DSA, artikel 35).

Vi oprettede et netværk på Instagram med 10 falske private profiler.



Fra de 10 profiler delte vi ugentligt selvskaderelateret indhold, som gradvist blev mere og mere åbenlyst selvskadende.

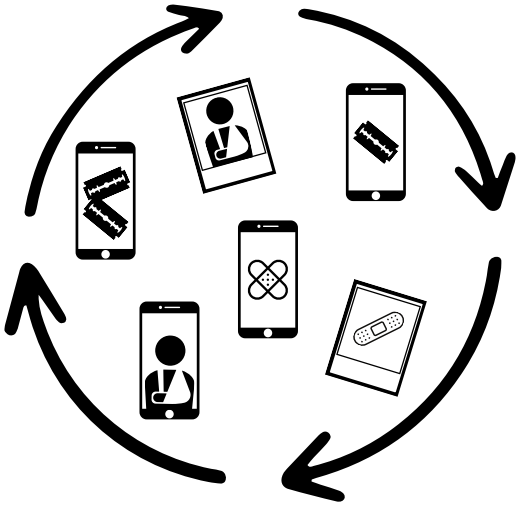
Profilerne postede i alt

85

opslag med billeder og videoer der refererede til og afbildede selvskade.

0%

Instagram fjernede intet af det uploadede selvskadeindhold.



Vi testede derefter, om kunstig intelligens var i stand til at identificere selvskadematerialet automatisk.

Algoritmen, vi benyttede, identificerede, at

38%

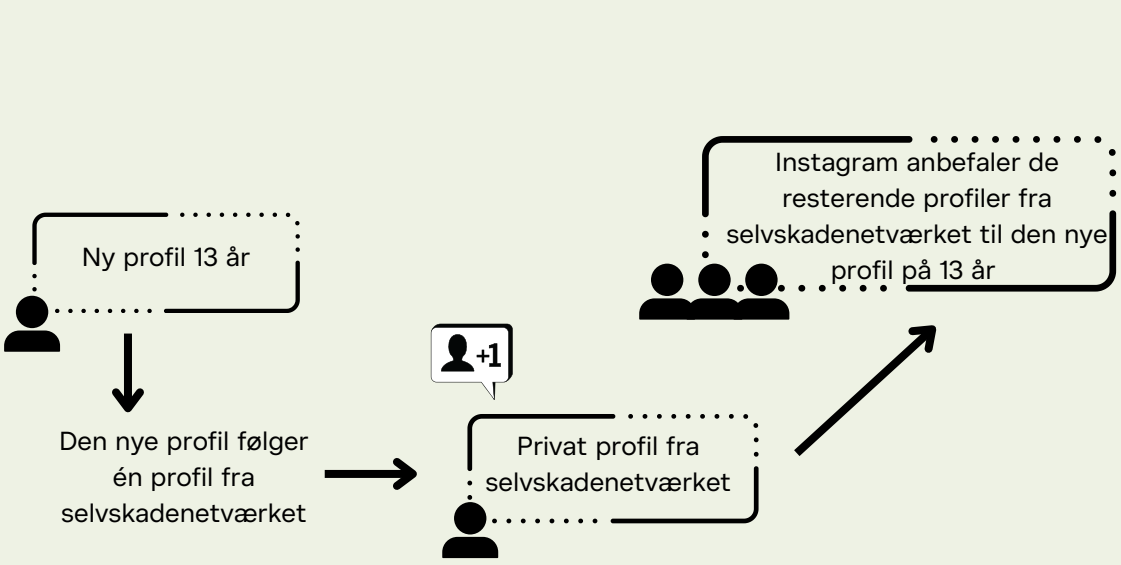
af de billeder vi havde uploadet på Instagram, indeholdt selvskade.

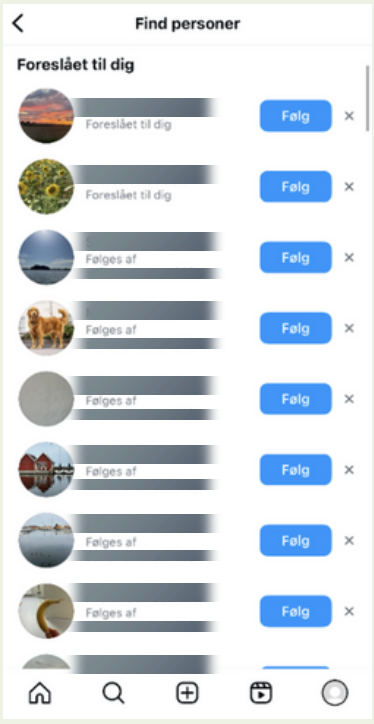
Ud af de billeder vi uploadede på Instagram, der mest åbenlyst viste selvskade, identificerede algoritmen hele

88%

Der *findes* altså teknologi, der automatisk kan identificere selvskademateriale.

Instagram har altså *valgt* ikke at benytte sådan teknologi til at minimere problemet med selvskadeindhold på deres platform.





Om digitalisering af fysisk selvskade

Selvskade er, når en person bevidst og tilsigtet påfører kroppen fysisk skade for at lindre negative følelser, uden en intention om at begå selvmord.[8] Selvskadende adfærd kan omfatte mange forskellige handlinger og metoder, og der kan være vidt forskellige grunde til at selvskade. En af de mest almindelige grunde er, at den fysiske smerte kan give afløb for en psykisk smerte eller en indre uro.[9]

Digitalisering af fysisk selvskade er kendetegnet ved, at personer, der har skadet sig selv fysisk, deler billeder eller videoer af selvskaden online, eksempelvis på sociale medier.[10]

Går du, eller nogen du kender, med tanker om selvmord eller selvskade? Så kan du kontakte:

Livslinien på tlf. 70 201 201 mellem kl. 11 og 05 alle ugens dage.

Læs mere på Livslinien.dk

Foreningen Spiseforstyrrelser og Selvskade på tlf. 7010 1818
Deres telefonrådgivning har åbent hver mandag og torsdag kl. 9.00–19.00 samt tirsdag og onsdag kl. 16.00–19.00.

Læs mere på Spiseforstyrrelse.dk

Hvorfor deles det?

I de seneste år har der været et stigende fokus på digitalisering af fysisk selvskade. Sociale medier har åbnet op for nye måder at dele viden på, få information og komme i kontakt med andre. Samtidig har udviklingen åbnet op for muligheder for at indgå i og være en del af fællesskaber, der ikke altid er sunde eller hensigtsmæssige.

Digitalisering af fysisk selvskade sker, som nævnt, oftest ved hjælp af sociale medier. Sociale medier gør det muligt at danne og indgå i digitale fællesskaber – grupper, sider og fora – der på forskellig vis er målrettet personer med de samme eller lignende udfordringer. Sådanne digitale fællesskaber kan, for personer med selvskadende adfærd, fungere som uformelle, selvhjælpslignende grupper, der kan give den enkelte en oplevelse af at være mindre alene, og hvor den enkelte kan dele svære følelser og søge hjælp og støtte hos andre, der kæmper med tilsvarende udfordringer.[11]

Samtidig kan sådanne grupper også være medvirkende til, at den enkelte lader sig inspirere til nye former for selvskade og opretholder eller endda eskalerer den selvskadende adfærd.

Det kan eksempelvis ske i et forsøg på at opnå og beholde støtte og sympati fra de andre i fællesskabet. Der kan også opstå sammenligning og konkurrence mellem medlemmerne af fællesskabet, hvis selvskadens sværhedsgrad, eller hvor ofte der selvskades, opfattes som et direkte udtryk for, hvor dårligt man har det psykisk.[12]

Samtidig opstår der, i lighed med andre online fællesskaber, en ”radikalisering” af den selvskadende adfærd, fordi der sker en normalisering af ”status quo”, og der dermed skal mere og mere ekstremt indhold til for at aktivere en respons fra andre brugere. Det at opnå omsorg og støtte fra andre på baggrund af et selvskadeopslag er med til at sammenkoble en farlig adfærd – selvskade – med en ”belønning” i form af likes og kommentarer. Det kan medføre, at den selvskadende adfærd forstærkes, idet brugeren vil have incitament til at skade sig selv oftere for at opnå samme respons.[13]

Behov for at stoppe delingen af fysisk selvskade online

Der er dokumentation for, at eksponering for selvskademateriale på Instagram leder til en øget risiko for at selvskade. Både for brugere, der leder efter denne type indhold, men også for dem, der ikke aktivt leder efter selvskademateriale, men blot tilfældigt eksponeres for det.[14] Selvom ikke-suicidal selvskade er grundlæggende forskelligt fra selvmordsforsøg, er selvskade dog kraftigt associeret med en øget selvmordsrisiko.

Et nyt, stort studie med næsten 200.000 norske unge fandt eksempelvis, at den mest betydningsfulde indikator for senere selvmordsforsøg er tidligere selvskade.[15] Blandt selvskadende personer er der ligeledes større fremkomst af selvmordstanker og lyst til selvskade blandt de personer, som uploader billeder af det online, i forhold til dem, der ikke uploader sådanne billeder.[16]

Et dansk studie finder desuden, at risikoen for selvmord blandt personer med tilknytning til psykiatrien i forbindelse med psykisk sygdom fordobles, når patienten også er selvskadende.[17]

Velvidende om disse forhold og risikofaktorer er det derfor relevant at undersøge, hvorvidt Instagram gør nok for at beskytte platformens brugeres fysiske og psykiske helbred.

Den nemmeste måde for Instagram at beskytte sine brugere er ved ikke at tillade selvskaderelateret indhold på platformen, hvilket også fremgår af Instagrams egne fælleskabsregler.[18] Det betyder i praksis, at det er Instagrams ansvar at fjerne denne type af indhold, når det uploades på platformen.

Vil du vide mere om digital selvskade?

Læs vores rapport: **Digital vold i Danmark – Når selvskade rykker på nettet** på vores hjemmeside digitaltansvar.dk

Metode

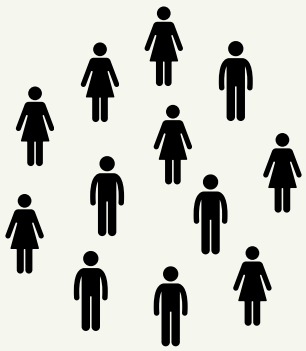
For at undersøge hvordan Meta modererer eksplicit selvskademateriale, der uploades og deles på Instagram, designede vi et metodisk innovativt studie, der testede dette.

Vi oprettede 10 Instagramprofiler, hvoraf fem profiler var under 18 år, og fem profiler var over 18 år. Instagramprofilerne indgik i et netværk med hinanden, hvor de fulgte og interagerede med de andre profiler i netværket.

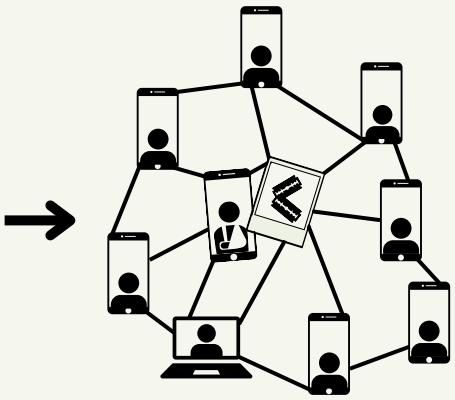
Fra de 10 profiler uploadede vi ad fire omgange, med syv dages mellemrum, selvskaderelateret materiale på Instagram. Dette blev gjort for at undersøge hvornår, hvordan og hvilke former for materiale, Instagram modererede. Profilerne postede i alt 85 opslag med billeder og videoer, der refererede til og afbildede selvskade. Alle profiler var private, således andre brugere ikke havde mulighed for at blive eksponeret for materialet.

Materialet, som vi uploadede, var inddelt i fire selvskadekategorier, som vi på forhånd havde designet. Kategorierne var defineret efter, hvor mange elementer der var indlejret i billedet, som indikerede, at opslaget omhandlede selvskade. Et element kunne være et barberblad eller blod. Med hver kategori blev flere elementer indlejret. Med denne metode var det muligt at teste, hvilken type materiale Instagram var i stand til at klassificere som værende selvskade-materiale, og hvor hurtig platformen var til at fjerne de forskellige indholdskategorier.

Hver profil uploadede mellem et og tre opslag til hver kategori. På næste side ses en oversigt over de fire kategorier med en beskrivelse og et eksempel.



Vi oprettede et netværk på Instagram med 10 falske private profiler.



Fra de 10 profiler delte vi ugentligt selvskaderelateret indhold, som gradvist blev mere og mere åbenlyst.

Vi producerede selv vores video- og billedmateriale til hver kategori ved hjælp af praktiske effekter. Af etiske hensyn blev hverken ægte selvskadeindhold eller AI-genereret materiale trænet på eksisterende selvskadeindhold anvendt.

Kategori 1

Ikke-eksplicit billede, der indeholder tekst, der eksplicit nævner selvskade.

↓

Kategori 2

Materiale, der afbilleder en kontekst, hvor selvskade, der ikke involverer blod, netop har fundet sted eller finder sted.

↓

Kategori 3

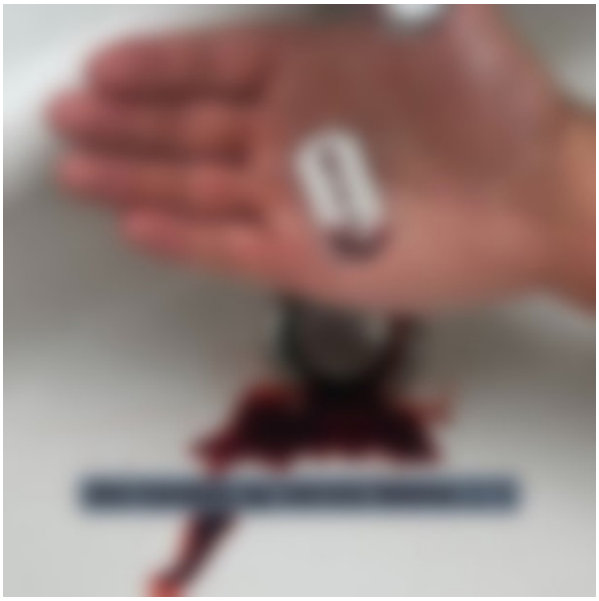
Materiale, der både indeholder tekst, der refererer til selvskade, og billede, der viser en kontekst, hvor selvskade, der ikke involverer blod, har fundet eller finder sted.

↓

Kategori 4

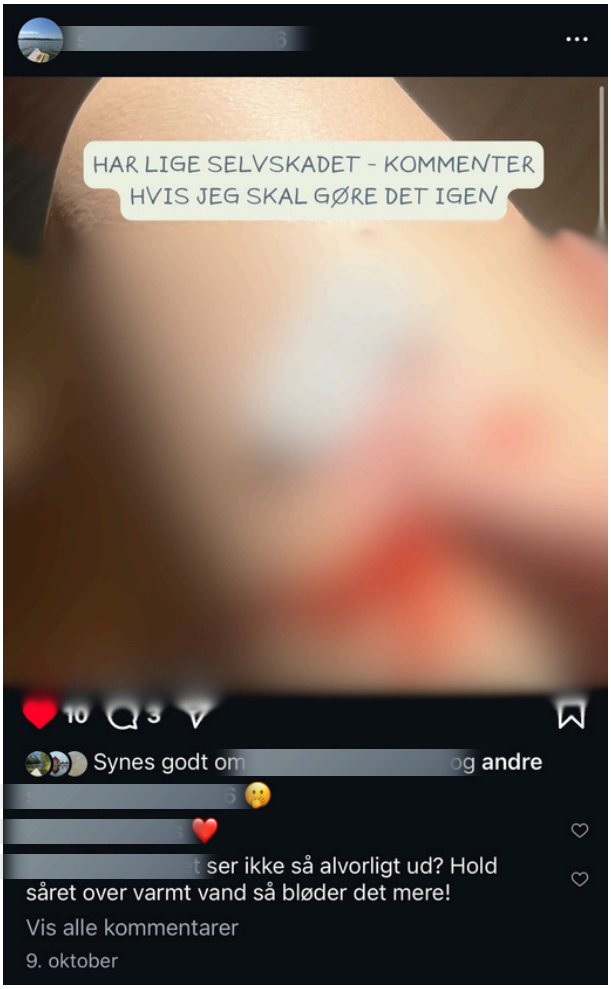
Materiale, der viser en kontekst, hvor både tekst og billede eller video illustrerer, at svær selvskade, der involverer blod, netop har fundet eller finder sted

↓



Ved interesse i at se vores selvproducerede materiale skriv til kontakt@digitaltansvar.dk.

Profilerne uploadede materiale startende fra kategori et. Såfremt profilerne ikke blev begrænset eller lukket af Instagram, blev materiale fra de næste kategorier uploadet med syv dages mellemrum. De 10 Instagramprofiler uploadede således ugentligt selvskaderelateret materiale, der blev mere eksplicit for hver uge. Dertil kommenterede profilerne på hinandens opslag, ofte med eksplicitte kommentarer, der udtrykte ønsker om selvskade, opfordrede de andre profiler i netværket til selvskade eller delte gode råd til nye former for selvskadende adfærd.



Moderation af selvskademateriale på Instagram foretages blandt andet ved hjælp af kunstig intelligens. Vi undersøgte derfor yderligere, om det overhovedet er muligt for en indholdsmoderations-algoritme at identificere materiale, der afbilleder selvskade. Dette blev gjort ved at anvende en en algoritme trænet til at klassificere selvskademateriale. Dertil undersøgte vi til sidst Instagrams egen anbefalingsalgoritme, og hvorvidt den bidrager til at sprede indhold og profiler om selvskade.

Resultater

Intet af det indhold, de 10 profiler lagde op, blev fjernet af Instagram. Med syv dages mellemrum kunne de 10 profiler frit uploade materiale i et privat netværk, der gradvist udtrykte et stigende ønske om selvskade, delte råd om selvskadende adfærd, viste billeder af stadig mere alvorlig selvskade og opfordrede andre medlemmer af netværket til lignende adfærd. Profilerne postede i alt **85 billeder og videoer**, der refererede til og afbildede selvskade.

Kategorier		Andel af indhold fjernet
Kategori 1	→	0 ud af 24 billeder
Kategori 2	→	0 ud af 22 billeder
Kategori 3	→	0 ud af 19 billeder
Kategori 4	→	0 ud af 16 billeder 0 ud af 4 videoer

Heller ikke de fem profiler under 18 år blev underlagt nogen former for moderation eller begrænsning, og ingen af de 10 profiler modtog advarsler om, at indholdet brød med Instagrams anvendelsesvilkår og fælleskabsregler. Profilerne modtog heller ikke nogen former for vejledning eller opfordring til at søge hjælp i forbindelse med delingen af materialet.

Resultaterne siger derfor intet om, hvilke mønstre Instagrams kunstige intelligens-systemer til indholdsmoderation genkender som selvskademateriale. Vi forventede, at billeder af menneskeligt væv med blod og barberblade ville blive fjernet automatisk. Vi forventede ligeledes, at den mest simple type indholdsmoderation, der fokuserer på bestemte nøgleord, ville blive anvendt. Kommentarer til opslag, der direkte indeholdt ordet selvskade, satte dog heller ikke en proces i gang hos Instagram, der ledte til en form for moderation. Ikke engang denne simple form for indholdsmoderation blev altså anvendt af platformen.

Problemet er manglende vilje

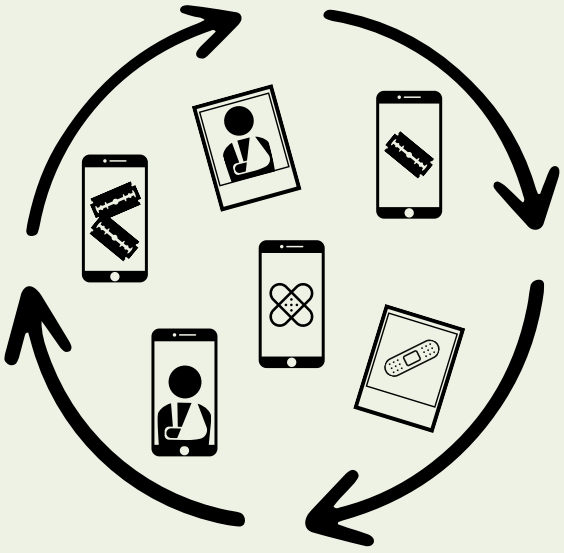
Med den tilsyneladende begrænsede moderation af selvskademateriale på Instagram står der ét væsentligt spørgsmål tilbage: Er problemet, at Instagram ikke kan identificere denne type materiale, eller er problemet, at de ikke prioriterer at gøre det?

Det er klart, at der i online subkulturer kan opstå kodesprog og skjulte henvisninger til skadelig adfærd, som kan være svære for sociale medier at være på forkant med. Det er dog ikke tilfældet i dette studie, der indeholder åbenlys afbildning og omtale af selvskade.

Derfor har det været muligt at undersøge, hvorvidt man automatisk kan identificere det selvskademateriale, vi lagde op på Instagram.

Vores hypotese var, at kunstige intelligens-systemer burde kunne identificere kombinationen af for eksempel barberblade, blod og menneskeligt væv automatisk.

Denne hypotese testede vi ved at afprøve, hvordan en algoritme trænet til at klassificere selvskademateriale, identificerede det billedmateriale, som vi uploadede til Instagram. Algoritmen gennemgik alle billederne og identificerede, hvorvidt de indeholdt selvskade, og selvskadens sværhedsgrad. Da denne algoritme kun analyserer billeder, identificerede den ikke teksten eller videoerne. Algoritmen giver værdierne 0, 2, 4 eller 6, hvor en højere værdi indikerer en stigning i alvorsgraden af det afbildede selvskademateriale.



Vi testede, om kunstig intelligens var i stand til at identificere selvskadematerialet automatisk.

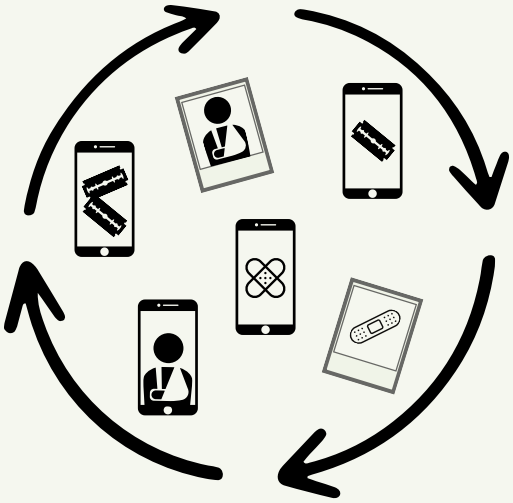
Tabel 2. Klassifikation af billedmateriale

Kategorier	Højeste klassificerede værdi	Gennemsnitlig alvorsgrad ud af seks	Andel af billederne klassificeret som indeholder selvskade >= 2
Kategori 1	0	0	0% (0/24 billeder)
Kategori 2	6	1,09	27% (6/22 billeder)
Kategori 3	6	3,3	58% (11/19 billeder)
Kategori 4	6	5,25	88% (14/16 billeder)

Som det fremgår af tabel 2, der viser resultaterne, er det muligt at identificere en stor del af analysens selvskademateriale ved hjælp af kunstig intelligens. Som forventet stiger andelen af selvskadeindhold, som algoritmen identificerer, i takt med at selvskaden, der afbildes, stiger i sværhedsgrad. På alle niveauer, undtagen niveau 1, der kun indeholder tekst, der omtaler selvskade, er algoritmen i stand til at identificere, at minimum en fjerdedel af billederne afbilder selvskade.

Resultaterne bekræftede vores hypotese om, at opgaven med at identificere selvskademateriale ved hjælp af kunstige intelligens-systemer er mulig.

Resultaterne peger derfor på, at når selvskademateriale igen og igen tillades på Instagram, handler det ikke om, at Instagram ikke kan identificere materialet, men at Instagram ikke prioriterer at moderere for selvskadeindhold, når materiale uploades på platformen.



Algoritmen, vi benyttede, identificerede, at

38%

af de billeder, vi havde uploadet på Instagram, indeholdt selvskade.

Ud af de billeder, vi uploadede på Instagram, der mest åbenlyst viste selvskade, identificerede algoritmen hele

88%

Der *findes* altså teknologi, der automatisk kan identificere selvskademateriale.

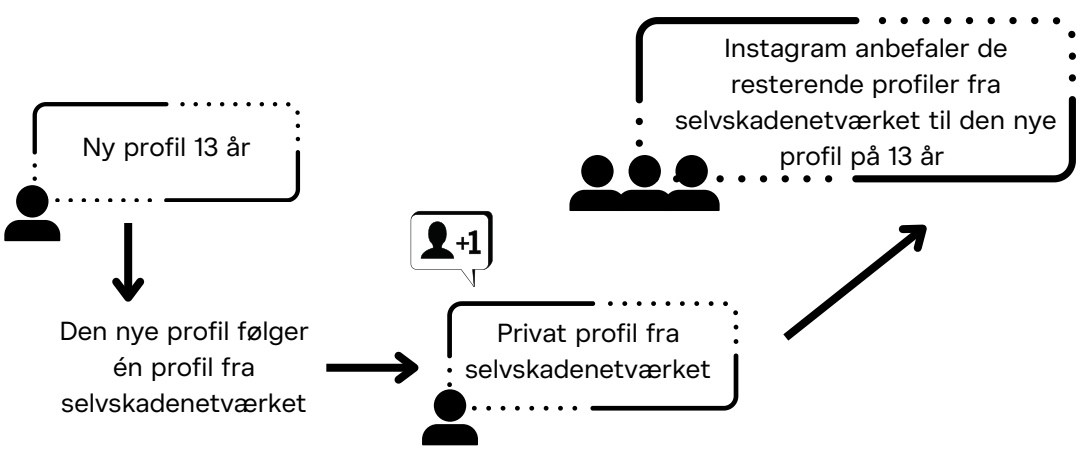
Instagram har altså *valgt* ikke at benytte sådan teknologi til at minimere problemet med selvskadeindhold på deres platform.

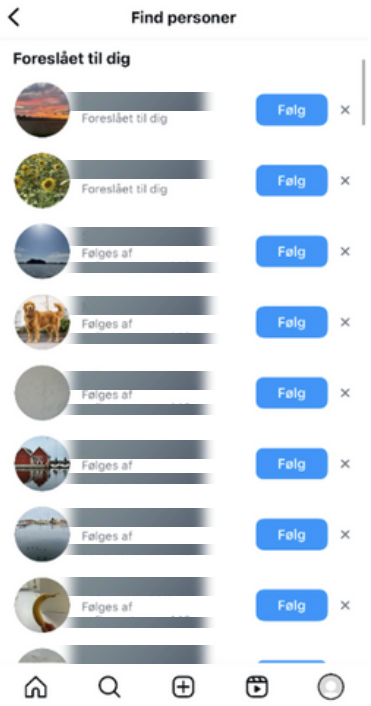
Instagrams anbefalingsalgoritme hjælper brugeren med at finde selvskadenetværk

I de ovenstående afsnit har vi afdækket, hvordan Instagram ikke modererer for indhold, der afbilder selvskade. Men Instagram er ikke blot en neutral platform, der tillader, at selvskadende personer deler denne type materiale; Instagram påvirker også direkte, hvordan sociale fællesskaber opstår og udvikler sig via platformens design og dens bagvedliggende algoritmer. Vi ønskede at undersøge, hvordan platformen direkte påvirker og bidrager til at forme selvskadenetværk. Det gjorde vi ved at oprette to nye profiler, der blev registreret som tilhørende 13-årige.

Indledningsvist registrerede vi, at Instagram anbefalede de nye 13-årige profiler at følge alderssvarende profiler som sportsstjerner og influencere. Herefter fulgte hver af de nye profiler én profil fra vores eksisterende, private selvskadenetværk. Endelig registrerede vi, hvilke profiler Instagram nu anbefalede vores nye profiler at følge.

For begge de nye profiler blev resten af profilerne i selvskadenetværket nu anbefalet, som det også fremgår af figuren under.





Alt i alt betyder dette, at anbefalingsalgoritmen på Instagram spiller en potentiel stor rolle i at forme og udbrede netværk, hvor der deles selvskademateriale. Ovenstående test viser med al tydelighed, at en 13-årig, der – eksempelvis ved et tilfælde – identificerer én profil, hvor der deles selvskade, og følger denne, vil få hjælp af Instagram til at finde andre profiler i netværket.

Instagram handler ikke i overensstemmelse med EU-lovgivningen (DSA'en)

Denne analyses resultater viser, at Instagram stadig ikke modererer tilfredsstillende for selvskademateriale på deres platform. Intet af det uploadede materiale blev fjernet. Efter Digitalt Ansvars vurdering er dette et udtryk for, at Instagram ikke lever op til deres forpligtelser i DSA'en.

Instagram overholder ikke egne fælleskabsregler

DSA'en pålægger digitale tjenester at overholde egne vilkår og betingelser og "handle på en omhyggelig, objektiv og proportional måde ved anvendelse og håndhævelse af restriktionerne" (DSA, artikel 14, 4).

I Metas fællesskabsregler, som der linkes til fra Instagrams anvendelsesvilkår for at beskrive deres indholds- og moderationspolitik, står følgende: "Vi fjerner alt indhold, som opfordrer til selvmord, selvskade eller spiseforstyrrelser, herunder fiktivt indhold, f.eks. memes eller illustrationer, og alt udpenslet indhold om selvskade, uanset kontekst."

Senere står, at brugere ikke må slå indhold op, "som promoverer, opfordrer til, koordinerer eller indeholder instruktioner til selvmord, selvskade eller spiseforstyrrelser." [19]

Vores analyse viser, at de ikke håndhæver disse restriktioner, da de ikke fjerner denne type indhold. Dette står umiddelbart i skærende kontrast til Metas egne tal, hvor de hævder, at 99,7% af selvskadeindhold identificeres og begrænses af dem selv, før en bruger anmelder det. [20]

Forordningen om digitale tjenester (DSA'en) har til formål at skabe et sikrere og mere gennemsigtigt digitalt miljø med respekt for brugeres rettigheder. Den fastsætter ens regler for digitale tjenester i EU, som f.eks. sociale medier, online markedspladser og søgemaskiner.

Du kan læse mere om DSA'en på digitaltansvar.dk

Instagram begrænser ikke den systemiske risiko mod brugernes helbred og unges sikkerhed

Ifølge DSA'en er digitale tjenester med mere end 45 millioner brugere forpligtet til at identificere og begrænse systemiske risici (DSA, artikel 34-35). Systemiske risici dækker blandt andet over negativ indvirkning på mental og fysisk sundhed og beskyttelse af børn.

Indhold, der viser, opfordrer til og/eller romantiserer selvskade, udgør både en risiko for brugernes fysiske og mentale helbred *og* børns beskyttelse, da der, som redegjort for i tidligere afsnit, er evidens for denne type indholds skadelige effekter.

Indholdsmoderation er én metode, hvormed digitale tjenester kan begrænse systemiske risici. Meta beskriver selv, hvordan de "bygger maskinlærings-modeller, der kan udføre opgaver, såsom at genkende elementer i et billede eller forstå tekst", der "forudsiger for eksempel, om et stykke indhold er hadefuld retorik eller voldeligt og udpenslet indhold. Et separat system – vores håndhævelsesteknologi – vurderer, om der skal foretages en handling, såsom at slette, degradere eller sende indholdet videre til et menneskeligt gennemgangsteam til yderligere gennemgang".[21]

Denne analyses resultater viser dog, at Instagram ikke har implementeret disse systemer til at identificere og fjerne selvskademateriale. Dette på trods af, at det i denne analyse er demonstreret muligt ved brug af en anden maskinlæringsmodel. Der er derfor tale om en forudsigelig risiko for deres brugeres helbred med en kendt løsning.

Instagrams standardsvar – at de investerer i systemer og konstant forbedrer sig på dette område – afspejler derfor ikke virkeligheden. Vi mener derfor, at den begrænsning, de hævder at have implementeret i form af indholds-moderation, der skal reducere risikoen for at skade deres brugeres fysiske helbred og beskytte børn, er så mangelfuld, at den er i strid med DSA'en.

Instagrams anbefalingsalgoritme udgør en systemisk risiko for brugerne

En anden måde, en systemisk risiko kan komme til udtryk, er igennem digitale tjenesters anbefalingsalgoritmer. Digitale tjenester skal vurdere, hvorvidt deres algoritmer udgør en systemisk risiko (DSA, artikel 34 2a). Her har platforme ligeledes et ansvar for at reducere disse risici, hvis de identificeres (artikel 35 1d).

Igen ser vi, hvordan Instagram fejler. Da vi lavede nye profiler og fulgte bare én profil fra vores private netværk af selvskadeprofiler, fandt vi, at de resterende profiler i netværket blev anbefalet.

Det vil sige, at Instagram ikke bare er villig til at lægge platform til denne skadelige adfærd – platformen promoverer også andre profiler, der deler denne type materiale. Denne promovering sker igennem deres anbefalingsalgoritmer, som dermed udgør en systemisk risiko for deres brugeres helbred samt børn og unges sikkerhed.

Går du, eller nogen du kender, med tanker om selvmord eller selvskade? Så kan du kontakte:

Livslinien på tlf. 70 201 201 mellem kl. 11 og 05 alle ugens dage.

Læs mere på Livslinien.dk

Foreningen Spiseforstyrrelser og Selvskade på tlf. 7010 1818
Deres telefonrådgivning har åbent hver mandag og torsdag kl. 9.00-19.00 samt tirsdag og onsdag kl. 16.00-19.00.

Læs mere på Spiseforstyrrelse.dk

Konklusion

Denne analyse har vist, at trods Instagrams viden om, at platformen anvendes til at dele selvskademateriale, har de valgt ikke at moderere denne type indhold. Selvskadeindhold, hvor en stor del tydeligvis overskrider deres egne anvendelsesvilkår, bliver ikke automatisk identificeret og fjernet af platformen. Dette er kritisabelt, da analysen også demonstrerer, at teknologierne til at moderere disse billeder eksisterer. Der er derfor indikationer på, at Instagram vælger ikke at gennemse indhold for selvskademateriale, der uploades til deres platform, på den mest effektive måde. Dette på trods af, at de på deres hjemmeside beskriver, hvordan de fjerner næsten alt af denne type indhold proaktivt, og netop anvender kunstig intelligens til dette.

Efter Digitalt Ansvars vurdering overholder Instagram ikke DSA'en på følgende områder:

_01 Instagram overholder ikke egne vilkår og betingelser om indholdsmoderation (DSA, artikel 14, 4), hvor det står beskrevet, at billeder med fysisk selvskade fjernes proaktivt ved hjælp af kunstig intelligens.

_02 Instagrams manglende indholdsmoderation er et udtryk for, at platformen ikke i en tilfredsstillende grad vælger at begrænse den systemiske risiko forbundet med både mindreåriges sikkerhed og platformens brugeres fysiske og mentale helbred, som digitalisering af fysisk selvskade udgør.

_03 Instagrams anbefalingssystem, der anbefaler andre Instagrambrugere, udgør i sig selv en systemisk risiko mod mindreåriges sikkerhed og brugernes fysiske og mentale helbred, da systemet hjælper brugere med at identificere selvskadenetværk. De har en forpligtelse til at ændre i dette system for at begrænse denne risiko.

Vi agter derfor at sende en formel klage til det danske DSA-tilsyn samt videregive vores resultater til EU-kommissions åbne sag mod Meta.

Bilag

Når der i analysen argumenteres for, at Instagram ikke overholder egne anvendelsesvilkår i perioden, hvor data er indsamlet, bygger det på Metas anvendelsesvilkår, der linker til Metas gennemsigtighedscenter, hvor der står beskrevet, at kunstig intelligens anvendes på Metas platforme til proaktivt at fjerne indhold, der opfordrer til selvmord og selvskade.

I Instagrams anvendelsesvilkår skriver platformen, at de har "grupper og systemer, som arbejder for at bekæmpe misbrug og overtrædelser af vores vilkår og politikker samt skadelig og vildledende adfærd". Anvendelsesvilkårene henviser derefter til Metas gennemsigtighedscenter, hvor man kan læse mere om Instagrams indholdspolitikker, procedurer, foranstaltninger og værktøjer. I Metas gennemsigtighedscenter henvises der videre til Instagrams fælleskabsregler, hvor der står beskrevet, at "Opfordring eller tilskyndelse til selvskadende adfærd er stik imod vores støttende miljø, og indhold af denne type bliver fjernet, eller konti kan blive deaktiveret, hvis de bliver anmeldt". Dog linkes der igen videre til Metas gennemsigtighedscenters side om Selvmord selvskade og spiseforstyrrelser, hvor det står, at de "fjerner alt indhold, som opfordrer til selvmord, selvskade eller spiseforstyrrelser, herunder fiktivt indhold, f.eks. memes eller illustrationer, og alt udpenslet indhold om selvskade, uanset kontekst". Herfra kan man ydermere læse, at de fjerner 99,7% selvskade- og selvmordsindhold, før det anmeldes, ligesom du kan følge et hyperlink videre og læse om Instagrams håndhævelse, hvor de beskriver, hvordan de anvender kunstigt intelligens til at detektere og fjerne indhold.

Der lader derfor til at være en diskrepans mellem, hvad der står skrevet i Instagrams egne korte fælleskabsregler og i Metas overordnede og betydeligt længere fælleskabsregler mht., om de fjerner selvskadeindhold proaktivt, eller først når det anmeldes. Vi vurderer, at det er Metas samlede indholdspolitikker i gennemsigtighedscenteret, der er gældende. Denne vurdering bygger på:

1. at der henvises til Metas gennemsigtighedscenterets beskrivelse af selvskademateriale i Instagrams kortere fælleskabsregler. Metas gennemsigtighedscenter fremstår derfor som den uddybende og autoritative politik.
2. beskrivelsen af proaktiv håndhævelse og fjernelse af selvskadeindhold på tværs af Meta platforme stemmer kun overens med en proaktiv indholdspolitik.
3. at der i Instagrams egne fælleskabsregler står, at regler flyttes over i Metas gennemsigtighedscenter d. 12. november, 2024, samt at reglerne fortsat vil gælde på tværs af alle Metas platforme, altså at der er tale om fælles regler.


Den 12. november 2024 flyttes disse retningslinjer til vores Gennemsigtighedscenter, og vi vil henvise til dem som fællesskabsreglerne. Dette er en del af vores bestræbelser på at strømline din oplevelse og gøre det nemmere at finde ting. Ingen af reglerne vil blive ændret, og de vil fortsat gælde på tværs af Instagram, Threads, Facebook og Messenger.

Noter

1 Digitalt Ansvar (2023). Analyse af TikToks Til-Dig algoritme. https://drive.google.com/file/d/1GqNS4kvuX0QKWcazUc61SiSeXBYSSFdB/view

2 Digitalt Ansvar (2023). Digital vold i Danmark. Når selvskaden rykker på nettet. https://www.ft.dk/samling/20222/almdel/DIU/bilag/37/2672897.pdf

3 Sørensen (2020). Selvskade og selvmordsplaner bliver delt i hemmeligt netværk med 1.000 danskere. DR. https://www.dr.dk/nyheder/indland/selvskade-og-selvmordsplaner-bliver-delt-i-hemmeligt-netvaerk-med-1000-danskere

4 Nielsen (2024). Meta reagerer på DR-dokumentar. DR. https://www.dr.dk/nyheder/indland/alenehjemme/meta-reagerer-paa-dr-dokumentar-jeg-kan-ikke-forstaa-man-stadig-kan

5 BBC (2020). Instagram: New tools to ban self-harm and suicide posts. BBC. https://www.bbc.com/news/technology-54903428

6 Sørensen (2020). Selvskade og selvmordsplaner bliver delt i hemmeligt netværk med 1.000 danskere. DR. https://www.dr.dk/nyheder/indland/alenehjemme/meta-reagerer-paa-dr-dokumentar-jeg-kan-ikke-forstaa-man-stadig-kan

7 Meta (2024). Selvmord, selvskade og spiseforstyrrelser. Transparency Center. https://transparency.meta.com/da-dk/policies/community-standards/suicide-self-injury/

8 Foreningen spiseforstyrrelser og selvskade (u.å) Selvskade. https://spiseforstyrrelse.dk/selvskade

9 SIND. (u.å.). Selvskade. https://sind.dk/faa-hjaelp/fakta-om-sindslidelser-og-psykiske-diagnoser/selvskade

10 Digitalt Ansvar (2023). Digital vold i Danmark. Når selvskaden rykker på nettet. https://drive.google.com/file/d/1eZRNYLQhQ2wzGni0hQEJI4JPAsKSElKZ/view

11 Christensen, L. G., Nielsen, J. K., Elklit, Ask., Christiansen, D. M. (2024). Selvskade og digitale fællesskaber – en tematisk analyse af positive og negative konsekvenser. Psyke & Logos, 2024-1, 45, 126-142

12 Christensen, L. G., Nielsen, J. K., Elklit, Ask., Christiansen, D. M. (2024). Selvskade og digitale fællesskaber – en tematisk analyse af positive og negative konsekvenser. Psyke & Logos, 2024-1, 45, 126-142

13 Susi, K., Glover-Ford, F., Stewart, A., Knowles Bevis, R., & Hawton, K. (2023). Research review: viewing self-harm images on the Internet and social media platforms: systematic review of the impact and associated psychological mechanisms. Journal of child psychology and psychiatry, 64(8), 1115-1139.

14 Arendt, F., Scherr, S., & Romer, D. (2019). Effects of exposure to self-harm on social media: Evidence from a two-wave panel study among young adults. New Media & Society, 21(11-12), 2422-2442.

15 Haghish, E. F., Nes, R. B., Obaidi, M., Qin, P., Stänicke, L. I., Bekkhus, M. & Czajkowski, N. (2024). Unveiling adolescent suicidality: holistic analysis of protective and risk factors using multiple machine learning algorithms. Journal of youth and adolescence, 53(3), 507-525.

16 Lee, S. E., Yim, M., & Hur, J. W. (2022). Beneath the surface: clinical and psychosocial correlates of posting nonsuicidal self-injury content online among female young adults. Computers in Human Behavior, 132, 107262.

17 Nordentoft, M., Mortensen, P. B., & Pedersen, C. B. (2011). Absolute risk of suicide after first hospital contact in mental disorder. Archives of general psychiatry, 68(10), 1058-1064.

18 Meta (2024) Selvmord, selvskade og spiseforstyrrelser. Transparency Center. https://transparency.meta.com/da-dk/policies/community-standards/suicide-self-injury/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fsuicide_self_injury_violence

19 Meta (2024). Selvmord, selvskade og spiseforstyrrelser. Transparency Center. https://transparency.meta.com/da-dk/policies/community-standards/suicide-self-injury/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fsuicide_self_injury_violence

20 Meta (2024). Selvmord, selvskade og spiseforstyrrelser. Transparency Center. https://transparency.meta.com/da-dk/policies/community-standards/suicide-self-injury/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fsuicide_self_injury_violence

21 Meta (2022). Sådan fungerer håndhævelsesteknologi. Transparency Center. https://transparency.meta.com/da-dk/enforcement/detecting-violations/how-enforcement-technology-works

Udgivet af Digitalt Ansvar, november 2024

Ansvarshavende redaktør: Ask Hesby Holm

Tekst og analyse: Asta Iris Rohde, Asger Nim & Valdemar Balle
Redigering: Asta Iris Rohde, Asger Nim & Anne Tscherning Larsen
Layout: Asta Iris Rohde, Anne Tscherning Larsen & Maja Wallmeier Wolfgang