

*Center for sociale medier, tech og demokrati
Slots- og Kulturstyrelsen
Hammerichsgade 14
1611 København V*

Høring over afgrænsning af "skadeligt indhold" og "skadelige funktioner" for børn og unge på sociale medier og andre digitale tjenester

Dataetisk Råd takker for den fremsendte høring. Rådet rådgiver regeringen og Folketinget om dataetiske problemstillinger knyttet til spørgsmålet om, hvordan anvendelse af ny teknologi kan ske med udgangspunkt i borgernes rettigheder, retssikkerhed og grundlæggende samfundsmæssige værdier. Spørgsmålet om hvordan man bør begrænse skadeligt indhold på digitale tjenester, særligt for at beskytte sårbare grupper som børn og unge, er således af central interesse for Dataetisk Råd.

Dataetisk Råd har i dette tilfælde valgt at afgive et høringssvar, fordi rådet har bemærkninger til definitionen af "skadeligt indhold", klassificeringen af visse typer indhold som skadelige, og til bruttolisten over potentielt skadelige funktioner.

De centrale pointer i Dataetisk Råds bemærkninger kan sammenfattes som følger:

- Dataetisk Råd finder det afgørende, at antage en præcis og velbegrundet definition af skadeligt indhold. Rådet foreslår følgende definition: "Skadeligt indhold er indhold som skader en person. Ved indhold forstås ethvert onlineindhold, herunder tekst, billeder, lyd, video og interaktive elementer. Ved skade forstås en reduktion af personens velfærd, for eksempel ved at eksponering for indholdet fører til fysisk smerte, psykisk lidelse, økonomisk tab, reduktion af selvværd, social eksklusion, eller tab af fysiske, sociale eller kognitive færdigheder."
- Dataetisk Råd vurderer, at risici for skade fortrinsvis er relevante i forbindelse med klassificering af indholdstyper som skadelige. Rådet vurderer, at det i denne forbindelse kan være relevant at vurdere i) risikoen for at en indholdstype gør skade, ii) omfanget og karakteren af den skade, som indholdet forårsager, iii) forekomsten af den pågældende indholdstype på digitale tjenester, iv) kumulative effekter af gentagen eksponering, samt v) hvordan risici skade varierer på tværs af relevante grupper. Rådet foreslår, at en indholdstype klassificeres som skadelig, når den gennemsnitlige brugers eksponering for indholdet sandsynligvis gør skade.

- Dataetisk Råd påpeger, at det forhold at indhold er skadeligt, eller at en indholdstype eller -funktion er skadelig, ikke i sig selv udgør et tilstrækkeligt grundlag for at begrænse brugeres adgang til sådant indhold. Det skyldes at det også vil være relevant at lægge vægt på hensyn, som kan trække i retning af at tillade skadeligt indhold, eksempelvis hensyn til indholdsproducenters ytringsfrihed og brugeres autonomi. Spørgsmålet om hvornår og hvordan skadeligt indhold bør reguleres eller modereres afhænger af hvordan disse hensyn etisk bør afvejes. Udformning af definitioner og fremadrettet anvendelse af den igangværende undersøgelse må således ske i lyset af en mere samlet dataetisk vurdering.

Rådet vil dog understrege, at vores bemærkninger i dette høringssvar er med forbehold for, at det er uklart hvilken anvendelse definitionen af skadeligt indhold, skadelige indholdstyper og -funktioner forventes at have, eksempelvis i et kommende lovarbejde, samt at der i det medsendte materiale refereres til afsnit i en rapport, som ikke er medsendt til høringen.

Definitionen af skadeligt indhold

Høringen over afgrænsning af "skadeligt indhold" og "skadelige funktioner" for børn og unge på sociale medier og andre digitale tjenester anfører i bilag 1 tre tilgange til at definere "skadeligt indhold": 1) en generel definition af skadeligt indhold, 2) en kategorisering af indholdstyper og 3) en hybrid tilgang, som kombinerer de to første tilgange. I bilag 1 nævnes desuden, at en kortlægning har vist at "der ikke eksisterer en universel generel definition af skadeligt indhold". Digitalt Ansvar anbefaler derfor den hybride tilgang, hvor skadeligt indhold både defineres og kategoriseres.

Rådet støtter anvendelsen af en kombination af en definition og en klassificering af indholdstyper, men bemærker i denne forbindelse, at definitionen af skadeligt indhold er central, idet det er rådets opfattelse, at det ikke er muligt at kategorisere konkrete indholdstyper som skadelige, uden at tage udgangspunkt i en definition, der gør det muligt at fastslå, om indhold som hører til denne type er eller ikke er skadeligt.

Høringen præsenterer på høringsbrevets side 2 og på side 2 i bilag 1 et forslag til en definition af skadeligt indhold, som "...indhold, der kan forårsage fysisk eller psykisk skade på en enkeltperson", idet det bemærkes, at "skade omfatter risiko for skade og potentiel skade", samt at "[v]ed indhold forstås ethvert onlineindhold, herunder tekst, billeder, lyd, video og interaktive elementer."

Rådet bemærker i denne forbindelse:

1. at det er afgørende for definitionen at præcisere og afgrænse begrebet "skade", idet Rådet lægger til grund, at formålet med definitionen er, at gøre det muligt at skelne mellem indhold som gør skade, og indhold som ikke gør skade. En definition af skadeligt indhold, som alene henviser til et ikke-nærmere defineret begreb om skade er i denne sammenhæng uhensigtsmæssigt cirkulær.
2. at en definition som klassificerer indhold som skadeligt når blot indholdet "kan forårsage skade", eller når der er "risiko for skade og potentiel skade" er upræcis og overinklusiv. Det vil eksempelvis betyde, at indhold som i en konkret situation ikke har gjort skade, ikke desto mindre skal klassificeres som skadeligt, hvis blot der fandtes en minimal risiko for, at indholdet kunne have gjort skade. Dataetisk Råd vurderer, at risici for skade vil snarere være relevant, når man vurderer hvordan indholdstyper og -funktioner bør reguleres, idet visse typer indhold vil have større og visse mindre risici for at være skadelige (se også bemærkninger nedenfor om klassificering af indholdstyper).
3. at forskningslitteraturen principielt definerer skade som reduktion af en persons velfærd, hvor velfærd forstås som de kvaliteter ved en persons liv, som gør livet værdifuldt for den pågældende person.¹

Med udgangspunkt i disse bemærkninger foreslår Dataetisk Råd følgende definition af skadeligt indhold: *"Skadeligt indhold er indhold som skader en person. Ved indhold forstås ethvert onlineindhold, herunder tekst, billeder, lyd, video og interaktive elementer. Ved skade forstås en reduktion af personens velfærd, for eksempel ved at eksponering for indholdet fører til fysisk smerte, psykisk lidelse, økonomisk tab, reduktion af selvværd, social eksklusion, eller tab af fysiske, sociale eller kognitive færdigheder."*

En sådan definition præciserer hvordan skadeligt indhold skal forstås på en måde, der gør det muligt principielt at skelne mellem skadeligt og uskadeligt indhold, og illustrerer definitionen med en række eksempler på hvordan en person kan blive skadet. Den udgør

¹ Se eksempelvis: Gardner, M. (2021). What Is Harming? In J. McMahan, T. Campbell, J. Goodrich, & K. Ramakrishnan (Eds.), *Principles and Persons: The Legacy of Derek Parfit* (pp. 381 – 395). Oxford University Press; Holtug, N. (2002). The Harm Principle. *Ethical Theory and Moral Practice*, 5(4), 357–389. <https://doi.org/10.1023/a:1021328520077>; Risberg, O., Johansson, J., & Carlson, E. (2021). Well-Being Counterfactualist Accounts of Harm and Benefit. *Australasian Journal of Philosophy*, 99(1), 164–174; Søbirk Petersen, T. (2014). Being Worse Off: But in Comparison with What? On the Baseline Problem of Harm and the Harm Principle. *Res Publica*, 20(2), 199–214; Purves, D. (2019). Harming as making worse off. *Philosophical Studies*, 176(10), 2629–2656.

derved et grundlag for at kunne klassificere bestemte indholdstyper og funktioner som skadelige.

Typer af skadeligt indhold

Med udgangspunkt i en definition af skadeligt indhold vil det være muligt, at klassificere visse indholdstyper som skadelige. Dette kan være relevant, hvis regeringen ønsker at regulere indholdstyper på digitale tjenester, eller at pålægge digitale tjenester at moderere indholdstyper, med henblik på at begrænse børn og unges eksponering for skadeligt indhold. Høringsbrevet anfører som eksempel, at typer af skadeligt indhold kan omfatte: "...indhold, der opfordrer til, promoverer eller giver instruktioner til selvmord, selvskade og ernærings- eller spiseforstyrrelser"

Dataetisk Råd bemærker i denne forbindelse at:

1. En "indholdstype" kan afgrænses på forskellig vis eksempelvis ved temaer, budskaber, formater eller effekter. Dataetisk Råd bemærker i denne forbindelse, at uanset hvordan en indholdstype afgrænses, vil der være få eller ingen indholdstyper, som nødvendigvis gør skade på alle brugere. Indholdstyper må derfor klassificeres som skadelige for så vidt *risikoen* for skade gør det sandsynligt, at brugere som eksponeres lider skade derved.
2. Det kan være relevant at vurdere både i) risikoen for at en indholdstype gør skade, ii) omfanget og karakteren af den skade, som indholdet forårsager, iii) forekomsten af den pågældende indholdstype på digitale tjenester, iv) kumulative effekter af gentagen eksponering, samt v) hvordan risici fordeles på tværs af relevante grupper, og hvordan skade varierer på tværs af relevante grupper, eksempelvis fordi visse grupper er mere sårbare end andre, når de udsættes for det pågældende indhold.
3. Det er afgørende for, hvordan definitionen af skadelige typer indhold udformes, hvilket formål en sådan definition skal tjene. Det forhold, at en indholdstype er skadelig, vil for nogle definitioner ikke i sig selv udgøre et tilstrækkeligt grundlag for at begrænse brugeres adgang til sådant indhold. Det skyldes at det også vil være relevant at lægge vægt på hensyn som kan trække i retning af at tillade skadeligt indhold, eksempelvis indholdsproducenters ytringsfrihed og brugeres autonomi. Spørgsmålet om hvornår og hvordan skadeligt indhold bør reguleres eller modereres afhænger af hvordan disse hensyn etisk bør afvejes. Det er således også muligt at anlægge en definition af skadelige indholdstyper, hvor en type indhold alene klassificeres som skadelig, når den forventede skade overstiger

en etisk tærskel, hvorefter brugere bør beskyttes mod at blive eksponeret for dette indhold.

De i pkt. 2. anførte overvejelser er relevante, eksempelvis fordi der kan være indholdstyper, som har lav risiko for at gøre skade, men er meget skadelige, når risikoen indfries. Et tænkt eksempel kunne være indhold som promoverer ekstremistiske konspirationsteorier, som langt de fleste brugere afviser, men som for enkelte brugere fører til radikalt afvigende og skadelig adfærd.

Tilsvarende kan der være indholdstyper, hvor det enkelte eksempel på indhold gør meget lidt skade, men hvor den kumulative effekt af massiv eksponering er væsentlig skade. Et tænkt eksempel kunne være indhold som forfægter urealistiske skønhedsidealiser, der isoleret set fremstår harmløst, men kumulativt fører til lavt selvværd, depression og selvskade for nogle brugere.

Kumulative effekter kan også være relevante, fordi der kan være visse typer indhold, hvor gentagen eksponering har en aftagende effekt, og andre hvor gentagen eksponering omvendt har stadigt stærkere effekt, samt indhold der for visse mængder eksponering ikke har nogen effekt, men hvor en effekt pludselig indtræder når den samlede eksponering når et vist omfang.

Vurderingen af risici og omfang af skade på tværs af grupper er relevant, fordi der kan være indholdstyper, som har meget lav risiko eller gør meget lidt skade for den gennemsnitlige bruger, men som for visse grupper har høj risiko, eller gør væsentlig skade.

I lyset af disse betragtninger foreslår Dataetisk Råd, at en indholdstype klassificeres som skadelig, når den gennemsnitlige bruger forventeligt lider skade ved at blive eksponeret for indholdet. Tilsvarende kan en indholdstype klassificeres som skadelig *for en bestemt gruppe* (herunder sårbare grupper), når det gennemsnitlige gruppemedlem forventeligt lider skade ved at blive eksponeret for indholdet.

Udvalgte potentielt skadelige funktioner

Af høringsbrevet og jf. bilag 2 forstår Rådet, at der ønskes input til, om en bruttoliste med "7 relevante potentielt skadelige funktioner" er dækkende til brug for det videre arbejde med undersøgelsen, herunder om der er behov for at inddrage andre typer af funktioner og i givet fald hvilke.

Dataetisk Råd har følgende bemærkninger:

1. Det fremgår ikke af høringsmaterialet, hvilket formål en bruttoliste over skadelige funktioner vil tjene, herunder om funktionerne på listen skal udelukkes eller ulovliggøres for børn og unge på digitale tjenester, om de i et eller andet omfang skal modereres, eller om formålet blot er at påpege, at de eksisterer og kan være skadelige. Ligesom det ovenfor er anført for definitionen af skadelige indholdstyper, er formålet imidlertid afgørende for, hvordan listen med skadelige funktioner bør udformes, inklusiv hvilke funktioner som bør optræde på listen. Ligesom for skadelige indholdstyper kan skadelige funktioner defineres som indholdsfunktioner der forventeligt gør skade på den gennemsnitlige bruger. Men ligesom for indholdstyper er det forhold, at en funktion er skadelig i denne forstand, ikke i sig selv tilstrækkeligt til, at en sådan funktion bør reguleres, idet denne konklusion forudsætter en afvejning af eventuelle konkurrerende hensyn. Derfor er det, ligesom for indholdstyper, også muligt at anlægge en definition, hvor en funktion alene defineres som skadelig, når den forventede skade overstiger en etisk tærskel, hvorefter funktionen bør reguleres.
2. Rådet noterer, at listen skal indeholde funktioner som " potentielt kan forårsage fysisk eller psykisk skade på en enkeltperson, og kan omfatte – men er ikke begrænset til bl.a. følgende 7 funktioner". Rådet vurderer, at de listede funktioner kan kategoriseres som skadelige jf. den under pkt.1 foreslåede definition. Det er imidlertid ikke klart, at de listede funktioner kan kategoriseres som skadelige, hvis man anlægger en definition hvorefter disse funktioner er *tilstrækkeligt* skadelige for den gennemsnitlige bruger til, at funktionerne bør reguleres. Hvorvidt dette er tilfældet vil afhænge af en konkret dataetisk afvejning for den enkelte funktion.
3. Høringssvarets bilag påpeger, at "Listen med indholdstyper ikke er udtømmende". Rådet bemærker i denne forbindelse, at det næppe vil være muligt at lave en udtømmende liste, og at det afhængigt af listens formål kan være relevant at overveje, hvordan listen skal holdes opdateret og relevant. Afhængigt af hvilket formål listen skal tjene, mener Dataetisk Råd at listen med fordel kan udvides. Den nuværende liste synes snæver i forhold til andre lignende internationale lister, herunder i UK Online Safety Act, World Economic Forums definitioner af "online harm", og i den videnskabelige litteratur².

² Se eksempelvis i Arora, Arnav, et al. "Detecting harmful content on online platforms: what platforms need vs. where research efforts go." ACM Computing Surveys 56.3 (2023): 1-17. & Scheuerman, Morgan Klaus, et al. "A framework of severity for harmful content online." Proceedings of the ACM on Human-Computer Interaction 5.CSCW2 (2021): 1-33.

Dataetisk konsekvensanalyse

Dataetisk Råd vil slutteligt benytte lejligheden til at minde om, at rådet generelt anbefaler, at der redegøres for de dataetiske konsekvenser af ændringer i bekendtgørelser og lovforslag. Dataetiske konsekvensanalyser sætter fokus på værdier og principper, som blandt andet velfærd og demokrati, værdighed, selvbestemmelse, lighed, gennemsigtighed, sikkerhed og privatliv. Dataetiske konsekvensanalyser kan således hjælpe med at bringe fordele, ulemper og utilsigtede konsekvenser ved lovforslag frem i lyset, og dermed bidrage til, at beslutninger tages på et mere kvalificeret grundlag. Dette gælder naturligvis i særlig grad ændringer og forslag, som angår persondata.

Dataetisk Råd kan i den forbindelse henvise til rådets værktøj '[Dataetik – Sådan gør du](#)', der operationaliserer identificeringen og stillingtagen til dataetiske dilemmaer.

Afsluttende bemærkninger

Dataetisk Råd anerkender og værdsætter Center for Medier og Techs initiativ til at undersøge en vigtig dataetisk problemstilling, i form af det skadelige indhold som brugere kan møde på digitale tjenester, og som særligt børn og unge kan være sårbare i mødet med. Rådet støtter udviklingen af en dansk definition af "skadeligt indhold", som kan have stor værdi, når vi som samfund skal forholde os til og tackle denne udfordring, og ser frem til at følge Centerets videre arbejde.

Dataetisk Råd står i den forbindelse til Centerets rådighed for uddybning og yderligere rådgivning.

På vegne af Dataetisk Råd og med venlig hilsen



Johan Busse

Formand