Nederlandse managementsamenvatting - Schimmige scheidslijnen:

Juridische, ethische, en praktische uitdagingen bij het detecteren en modereren van terroristische, extremistische en impliciete extremistische content, met respect voor de vrijheid van meningsuiting

Bibi van Ginkel, Tanya Mehra, Merlina Herbach, Julian Lanchès, en Yael Boerma

International Centre for Counter-Terrorism

Nederlandse managementsamenvatting - Schimmige scheidslijnen:

Juridische, ethische, en praktische uitdagingen bij het detecteren en modereren van terroristische, extremistische en impliciete extremistische content, met respect voor de vrijheid van meningsuiting

Bibi van Ginkel, Tanya Mehra, Merlina Herbach, Julian Lanchès, en Yael Boerma

ICCT Report

November 2025

International Centre for Counter-Terrorism

About ICCT

The International Centre for Counter-Terrorism (ICCT) is an independent think and do tank providing multidisciplinary policy advice and practical, solution-oriented implementation support on prevention and the rule of law, two vital pillars of effective counter-terrorism.

ICCT's work focuses on themes at the intersection of countering violent extremism and criminal justice sector responses, as well as human rights-related aspects of counter-terrorism. The major project areas concern countering violent extremism, rule of law, foreign fighters, country and regional analysis, rehabilitation, civil society engagement and victims' voices. Functioning as a nucleus within the international counter-terrorism network, ICCT connects experts, policymakers, civil society actors and practitioners from different fields by providing a platform for productive collaboration, practical analysis, and exchange of experiences and expertise, with the ultimate aim of identifying innovative and comprehensive approaches to preventing and countering terrorism.

Licensing and Distribution

Photocredit: Pravit/Adobe Stock

Inleiding en Achtergrond

In juni 2024 gaf het Dreigingsbeeld Terrorisme Nederland (DTN) een scherpe waarschuwing af: terroristische en extremistische actoren blijven onlineplatforms misbruiken om propaganda te verspreiden, activiteiten te coördineren en tot geweld aan te zetten. Deze content varieert van openlijk terroristisch materiaal tot zogenoemde *'borderline'*-content - materiaal dat niet duidelijk binnen de wettelijke definities van terroristische of illegale content valt, maar desalniettemin een erosief effect heeft op democratische waarden en sociale cohesie. Daarom wordt dit ook wel aangeduid als *'awful but lawful'* ('afschuwelijk maar wettelijk toegestaan').

De online verspreiding van schadelijke content is niet nieuw, maar de reikwijdte, snelheid en verfijning ervan zijn door technologische innovaties drastisch toegenomen. De opkomst van generatieve kunstmatige intelligentie (AI) en grote taalmodellen (LLM's) stelt extremistische actoren in staat om content te creëren en te verspreiden op ongekende schaal, snelheid en precisie, vaak in meerdere talen tegelijk. Jihadistische groeperingen zoals ISIS en Al-Qaida, evenals rechts-extremistische groeperingen in Europa en Noord-Amerika, maken reeds gebruik van deze technologieën in hun propagandacampagnes.

De belangen zijn groot. Dergelijke content kan haat aanwakkeren, terroristisch geweld normaliseren en maatschappelijke polarisatie verdiepen. Jongeren, die hun informatie voornamelijk via sociale media verkrijgen, zijn bijzonder kwetsbaar voor online radicalisering. Extremistische actoren misbruiken niet alleen reguliere sociale mediaplatforms, maar ook gamingplatforms, streamingwebsites en *file sharing* omgevingen. De verschuiving van deze content van obscure, moeilijk toegankelijke uithoeken van het internet tien jaar geleden naar openbaar toegankelijke platforms vandaag de dag benadrukt de urgentie van dit probleem.

Tegelijkertijd roept dit vraagstuk complexe dilemma's op die op het snijvlak van veiligheid, technologie en fundamentele rechten liggen. Schadelijke content wordt vaak verhuld met humor, ironie of satire, waardoor de grens tussen vrijheid van meningsuiting en aanzetten tot geweld vervaagt. Te brede contentmoderatie kan legitiem debat onderdrukken, terwijl ontoereikende moderatie samenlevingen blootstelt aan manipulatie en radicalisering.

De rol van de technologiesector is cruciaal bij het aanpakken van de verspreiding van schadelijke content, maar hun aanpak is niet eensluidend en staat steeds vaker ter discussie. Hoewel platforms een grote verantwoordelijkheid dragen voor detectie en moderatie, blijft hun samenwerking met onafhankelijk onderzoek en publieke instellingen beperkt. Deze terughoudendheid weerspiegelt bredere zorgen dat grote technologiebedrijven prioriteit geven aan winstgevendheid van bedrijfsvoering ten koste van maatschappelijke verantwoordelijkheid. Ze investeren fors in geautomatiseerde detectiesystemen, terwijl teams van menselijke moderatoren tegelijkertijd worden ingekrompen. Dergelijke trends dreigen zowel de kwaliteit als de legitimiteit van moderatie te ondermijnen, aangezien algoritmische hulpmiddelen op zichzelf vaak ongeschikt zijn om de nuance en context van impliciete extremistische content te herkennen.

Tegen deze achtergrond gaf het Wetenschappelijk Onderzoek- en Datacentrum (WODC) op verzoek van de Nationaal Coördinator Terrorismebestrijding en Veiligheid (NCTV) opdracht tot een studie - uitgevoerd door ICCT - naar de haalbaarheid van het opstellen van een duidingskader dat een betrouwbare detectie en moderatie van extremistische en terroristische, online content mogelijk maakt, zonder dat dit ten koste gaat van de vrijheid van meningsuiting. Onderdeel van de categorie problematische content, is de zogenoemde 'borderline content', die vaak een impliciet karakter heeft, en daardoor niet altijd gemakkelijk te detecteren is.

De Maatschappelijke Uitdaging

De grote hoeveelheid schadelijke content die online beschikbaar is stelt democratische rechtsstaten voor grote uitdagingen. Dergelijke content erodeert immers pluralisme, vertrouwen en sociale cohesie, waarden die het fundament vormen van een democratische rechtsstaat. Terroristische propaganda, extremistische narratieven en meer impliciete vormen van haatdragende of polariserende taal richten zich niet alleen op individuen of groepen; ze beogen democratische instituties te destabiliseren door geweld te normaliseren, polarisatie te voeden en het vertrouwen in het vermogen van de staat om burgers te beschermen te ondermijnen. Zonder toezicht kan dergelijke content gevoelens van onvrede versterken, maatschappelijke breuklijnen verdiepen en de principes van vrij en open debat die essentieel zijn in een democratische samenleving ondermijnen. Onlineplatforms zijn de hedendaagse *agora,* het dorpsplein van het moderne publieke leven. Ze bieden ruimte voor politiek debat, culturele uitwisseling en sociale interacties. Echter, deze ruimtes worden ook gebruikt door extremistische en terroristische actoren die communicatiemiddelen misbruiken om ideologische agenda's te bevorderen.

Deze managementsamenvatting bundelt de achtergrond, onderzoeksvragen, bevindingen, uitdagingen en aanbevelingen van dat onderzoek. Ze biedt een kritische reflectie op de mogelijkheden en beperkingen van contentdetectiekaders en geeft concrete aanbevelingen aan beleidsmakers, online dienstverleners en andere belanghebbenden. Vooraf definiëren we de reikwijdte van het onderzoek en lichten we de gebruikte methodologie toe.

Reikwijdte

Gezien de onduidelijkheid rond bestaande definities - met name het concept *borderline content* - achtte het onderzoeksteam het noodzakelijk om af te wijken van de veel gebruikte terminologie. Voor de doeleinden van dit onderzoek en om de reikwijdte te beperken, hanteren we de categorieën *terroristische content*, *illegale content* en *impliciete extremistische content* die schadelijk is in de context van gewelddadig extremisme en terrorisme. Deze keuze is vooral relevant voor de haalbaarheidsvraag aangaande een duidingskader. Immers, vaststellen of content valt onder de vrijheid van meningsuiting vereist het gebruik van duidelijke juridische definities. Beperkingen op uitingen zijn alleen legitiem wanneer ze in de wet zijn verankerd en voldoen aan de criteria van proportionaliteit, noodzakelijkheid en adequaatheid. Omdat noch 'extremistische content' noch 'borderline content' juridisch is gedefinieerd, bestaat het risico dat verschillende belanghebbenden dit uiteenlopend interpreteren. Daarom kozen we voor de term *impliciete extremistische content* om materiaal te beschrijven dat op het eerste gezicht niet als terroristisch of illegaal herkenbaar is vanwege de verhullende aard, maar toch schadelijk is en potentieel bevorderlijk is voor radicalisering.

Methodologie

Dit onderzoek maakte gebruik van een *mixed-methods* aanpak, waarbij literatuuronderzoek, semigestructureerde interviews, een expertbijeenkomst en kwalitatieve inhoudsanalyse werden gecombineerd. Het literatuuronderzoek omvatte academische literatuur, beleidsrapporten, wetgeving, jurisprudentie en servicevoorwaarden van platforms gepubliceerd voor juni 2025. Dit vormde de basis voor het identificeren van indicatoren die relevant zijn voor een duidingskader. Semigestructureerde interviews werden afgenomen met relevante belanghebbenden - waaronder overheidsinstanties, wetshandhavingsinstanties, EU-organen en ngo's - al weigerden platforms mee te werken. Het gebrek aan bereidheid mee te werken aan dit onderzoek, illustreert de aanhoudende zorgen over het algehele gebrek aan medewerking en transparantie in de sector. De inzichten werden aangevuld met een rondetafelgesprek met experts en praktijkdeskundigen, gericht op definitie-uitdagingen, detectie- en moderatiemethoden en de haalbaarheid van een duidingskader.

Om de operationalisering van geïdentificeerde indicatoren te toetsen, ontwikkelde het onderzoeksteam een codeboek dat werd getest door middel van een kwalitatieve inhoudsanalyse van online data, verzameld via *open source intelligence* (OSINT) onderzoek. Voor deze vorm van een kwalitatieve inhoudsanalyse gebruikten wij een contactloze en geanonimiseerde scrape methode om online content gepost op verschillende accounts te verzamelen. Voorafgaand aan het OSINT onderzoek, is een *data protection impact assessment* uitgevoerd en bovendien werden strikte ethische waarborgen in acht genomen. Het doel van het OSINT onderzoek was om een pilot haalbaarheidsstudie uit te kunnen voeren van een selectie indicatoren die gebruikt kunnen worden bij het identificeren en kwalificeren van content in een duidingskader.

De geselecteerde content voor deze exercitie werd dus niet gebruikt om inzicht te krijgen in de mate, hoeveelheid of soorten van extremistische en terroristische content op de geselecteerde platforms. Dergelijke vragen werden beantwoord op basis van het literatuuronderzoek. Voor het OSINT onderzoek werden drie platforms geselecteerd - Instagram, TikTok en Reddit - op basis van criteria zoals relevantie voor de Nederlandse context, toegankelijkheid voor open bronnenanalyse, diversiteit van gebruikersdemografie en variatie in moderatiepraktijken (bijv. AI-dominant op TikTok en Instagram, hybride mens-automatische aanpak op Reddit).

Er werd content verzameld met betrekking tot twee nationaal significante gebeurtenissen: de Amsterdamse rellen naar aanleiding van een confrontatie tussen voetbalsupporters van Maccabi Tel Aviv en een groep mensen met sterke sentimenten rond het conflict tussen Israël en Gaza, die bovendien uiting gaven aan antisemitische gevoelens en Maccabi-fans opjaagden en aanvielen (content geplaatst in november 2024) en de "*White Lives Matter*"-projectie op de Erasmusbrug tijdens de Oud-en-nieuwviering (content geplaatst in januari 2023 met betrekking tot het incident van de projectie en december 2024 - januari 2025 met betrekking tot de rechtszaak). Uiteindelijk werden er negen accounts geselecteerd op de drie platforms, die zowel rechts-extremistisch als jihadistisch geïnspireerde narratieven weerspiegelden. Alle berichten die door deze accounts werden gepubliceerd binnen de observatieperiode werden handmatig gescrapet, geanonimiseerd en veilig opgeslagen. Vervolgens werden de berichten gecodeerd met behulp van een pilot codeboek, dat indicatoren testte in de drie categorieën - terroristische, illegale en impliciete extremistische content - wat empirische input leverde over de haalbaarheid van een duidingskader.

Typen Schadelijke Content

Schadelijke online content manifesteert zich in verschillende overlappende categorieën. Voor dit rapport richten we ons op de categorieën die relevant zijn voor terrorisme en radicalisering tot (gewelddadig) extremisme:

- **Terroristische content** wordt gedefinieerd in overeenstemming met Verordening (EU) 2021/784, waarin staat dat terroristische content elk materiaal omvat dat: (i) iemand aanzet of oproept tot het plegen van terroristische strafbare feiten of daaraan bij te dragen, (ii) oproept tot deelname aan activiteiten van een terroristische groepering, (iii) terroristische activiteiten verheerlijkt, onder meer door materiaal te delen dat terroristische aanslagen afbeeldt, of (iv) instructies geeft voor het maken of gebruiken van explosieven, vuurwapens of andere wapens, waaronder chemische, biologische, radiologische of nucleaire stoffen. Daarbij worden terroristische strafbare feiten gedefinieerd overeenkomstig artikel 3 van Richtlijn (EU) 2017/541.

- **Illegale content** verwijst naar online content die in strijd is met nationale of Europese wetgeving. Dit omvat content die op zichzelf illegaal is, evenals content die inbreuk maakt op consumentenbeschermingswetten of een schending vormt van intellectuele eigendomsrechten. Voor de reikwijdte van dit onderzoek richten we ons uitsluitend op

illegale content in de context van terrorisme en gewelddadig extremisme. Dit kan onder meer haatzaaiende uitlatingen omvatten en online content die bijdraagt aan polarisatie en radicalisering.

- **Impliciete extremistische content die schadelijk is:**
  - **'Extremistisch'** verwijst naar content die narratieven verspreidt die groepen of personen buitensluiten of haatdragend van aard zijn, en die kunnen bijdragen aan radicalisering richting terrorisme en (gewelddadig) extremisme.
  - **'Impliciet'** verwijst naar het feit dat de betekenis verhuld is. Wanneer dit opzettelijk gebeurt, is het doel om de illegaliteit, onrechtmatigheid of schadelijkheid van de content te verhullen.
  - **'Schadelijk'** verwijst naar het feit dat de content ernstige schade kan toebrengen aan een individu, een groep mensen, instellingen of de democratische rechtsorde, en dat deze niet wordt beschermd door internationale mensenrechtenrechten.

De categorie impliciete extremistische content is in het bijzonder problematisch. De meerduidigheid van deze content beschermt tegen directe juridische sancties en maakt het eveneens mogelijk verdeeldheid te zaaien en extremistische wereldbeelden te versterken. Bovendien stemmen extremistische actoren hun berichtgeving bewust af om binnen de grijze zone te blijven, wat ervoor zorgt dat content aan moderatie ontsnapt en toch radicaliserend werkt.

Verhulling en Aanpassing

Extremistische actoren gebruiken verhullingsstrategieën om detectie te omzeilen. *Dog whistles*, gecodeerde emoji's en historische of culturele verwijzingen die alleen voor ingewijden begrijpelijk zijn, worden vaak ingezet. Humor en ironie, vooral via memes, dienen zowel als retorisch schild als rekruteringsmiddel. Ze normaliseren extremistische ideeën en weerleggen tegelijk externe kritiek.

Deze actoren zijn wendbaar en adaptief. Naarmate platforms strengere moderatie toepassen op openlijk terroristische content, verschuiven extremisten naar meer impliciete vormen van content waarbij ze zorgvuldig hun discours zo formuleren om het *'awful but lawful'* te laten lijken. Opkomende technologieën versterken deze trend: generatieve AI faciliteert de productie van geavanceerde teksten, beelden en video's, terwijl *deepfake*-technologieën en interactieve game-omgevingen nieuwe mogelijkheden voor verspreiding bieden.

Maatschappelijke Gevolgen

De gevolgen van ongecontroleerde schadelijke content zijn aanzienlijk:

- **Radicalisering**: blootstelling aan extremistische narratieven online is een aantoonbare factor van gewelddadige radicalisering, vooral onder jongeren.
- **Normalisering van geweld**: herhaaldelijke blootstelling aan (gewelddadige) extremistische retoriek vermindert de maatschappelijke weerstand tegen geweld, wat extremistische wereldbeelden inbedt in het reguliere discours.
- **Polarisatie**: schadelijke content verdiept problematische maatschappelijke verdeeldheid, erodeert vertrouwen in instituties en voedt vijandigheid tussen gemeenschappen.
- **Democratische weerbaarheid**: manipulatie van online discours ondermijnt het open democratische debat, wat de ruimte voor pluralisme en constructief meningsverschil vernauwt.

De maatschappelijke uitdaging is dus tweeledig: voorkomen dat onlineplatforms worden misbruikt voor extremistische doeleinden, en tegelijkertijd fundamentele vrijheden, waaronder de vrijheid van meningsuiting, waarborgen.

Het regelgevende en institutionele landschap

Europese Kaders

Als antwoord op deze dreigingen heeft de Europese Unie een gelaagd regelgevend kader ontwikkeld. Twee instrumenten springen eruit:

- **Verordening inzake verspreiding van terroristische online-inhoud (TCO, Verordening EU 2021/784):** verplicht hostingdiensten (HSP's) om terroristische content binnen één uur na melding door bevoegde autoriteiten te verwijderen. De verordening introduceert transparantievereisten, gebruikersmelding en gedifferentieerde verplichtingen afhankelijk van het blootstellingsniveau. Critici wijzen op de onevenredige last voor kleinere platforms die onvoldoende middelen hebben om te voldoen aan de verplichtingen.
- **Digitaledienstenverordening (DSA, Verordening EU 2022/2065)**: verbreedt de regelgevende reikwijdte aanzienlijk om tegen alle illegale content op te treden en legt zeer grote onlineplatforms (VLOP's) en onlinezoekmachines (VLOSE's) verplichtingen op om systemische risicobeoordelingen uit te voeren. De DSA beoogt transparantie te vergroten door jaarverslagen te vereisen, een DSA-transparantieregister te creëren en onderzoekers toegang te geven tot platformgegevens.

Gezamenlijk markeren deze kaders een verschuiving van vrijwillige samenwerking met platforms naar een regelgevingsmodel dat streeft naar harmonisatie, verantwoordelijkheid en minimale waarborgen voor de mensenrechten. Toch blijven er uitdagingen voor de implementatie, met name rond proportionaliteit, capaciteitsbeperkingen van kleinere aanbieders en de bescherming van fundamentele rechten.

Platformbestuur

Naast uitvoering geven aan de juridische verplichtingen, handhaven platforms ook hun eigen gebruiksvoorwaarden (ToS) en communityrichtlijnen. Deze regels reiken vaak verder dan wettelijke verplichtingen en omvatten bredere categorieën van schadelijke content. Hoewel deze proactieve houding kan bijdragen aan het vermijden van bestuurlijke boetes, roept ze zorgen op over het feit dat private actoren in de praktijk de grenzen van online vrijheid van meningsuiting bepalen, zonder duidelijk democratisch toezicht.

Een gebrek aan transparantie verergert dit probleem. Gebruikers begrijpen vaak niet waarom content wordt verwijderd, terwijl onderzoekers en toezichthouders obstakels ondervinden bij het verkrijgen van moderatiegegevens. De ondoorzichtigheid over handhavingsbesluiten vanuit de ToS ondermijnt vertrouwen en verantwoordelijkheid.

Dimensie van Fundamentele Rechten

Het recht op vrijheid van meningsuiting staat centraal in dit debat. Contentmoderatie omvat onvermijdelijk normatieve oordelen over wat toelaatbaar is. Wanneer platforms het zekere voor het onzekere nemen om boetes te voorkomen, lopen zij het risico legitieme kritiek, satire of dissidentie te verwijderen, met onevenredige effecten op gemarginaliseerde of minderheidsgroepen. Daar tegenover staat de ontoereikende moderatie die ertoe kan leiden

dat schadelijke narratieven ongestoord gedijen. Uiteindelijk is het laatste oordeel aan de rechter bij geschillen, maar rechtszaken verlopen traag en zijn gering, en bieden weinig houvast in de tussentijd.

De uitdaging op het gebied van beleid- en regelgeving gaat dus niet alleen over handhaving naleving, maar ook over het inbedden van waarborgen voor fundamentele rechten in moderatiepraktijken.

Onderzoeksvragen

Het WODC heeft, op verzoek van de NCTV, een onderzoek laten uitvoeren naar de vraag of er een betrouwbaar duidingskader kan worden ontwikkeld om platforms te helpen bij het identificeren en modereren van terroristische, illegale en impliciete extremistische content. Er stonden zeven kernvragen centraal in het onderzoek:

1. Welke kenmerken of combinatie van kenmerken in online content bepalen/bepaalt of er sprake is van extremistische of terroristische content op onlineplatforms?
2. Op welke wijze en in welke mate zijn deze kenmerken, al dan niet in combinatie met elkaar, ook te detecteren en te identificeren in de content op onlineplatforms?
3. Wat kan er worden gezegd over de betrouwbaarheid van de detectiemogelijkheden voor extremistische en terroristische content op onlineplatforms?
4. Hoe verhoudt de betrouwbaarheid van de detectiemogelijkheden voor de verschillende typen van schadelijke content op onlineplatforms (terroristisch, (evident) extremistisch en borderline) zich tot het risico op onterechte moderatiebeslissingen door onlineplatforms waarbij het fundamentele recht op vrijheid van meningsuiting wordt geschonden?
5. Is het met de huidige kennis en stand van de techniek mogelijk om een valide en betrouwbaar duidingskader op te stellen voor de detectie en identificatie van zowel de expliciete extremistische en terroristische content als de meer impliciete borderline content op onlineplatforms, zonder dat onterecht het fundamentele recht op de vrijheid van meningsuiting geweld wordt aangedaan?
6. Onder welke condities kan de toepassing van een duidingskader voor extremistische en terroristische online content bijdragen aan de vermindering van zowel de verspreiding van deze schadelijke content als de potentiële radicalisering van internetgebruikers?
7. Indien een valide en betrouwbaar duidingskader niet mogelijk geacht wordt: wat zijn de juridische, technische en eventueel andere knelpunten/barrières die het opstellen van een duidingskader voor extremistische online content in de weg staan? Aan welke voorwaarden zou moeten worden voldaan om deze knelpunten/barrières weg te nemen en wat is de uitvoerbaarheid hiervan?

Deze vragen vormden de analytische lens voor het beoordelen van haalbaarheid, betrouwbaarheid en ethische wenselijkheid.

Belangrijkste Bevindingen

Op basis van het literatuuronderzoek heeft het team essentiële indicatoren om terroristische, illegale en impliciete extremistische online content te kwalificeren geïdentificeerd (onderzoeksvragen 1 en 2), en deze gebruikt om een pilot codeboek te ontwikkelen. Het doel van het onderzoek was echter niet om een volledig duidingskader te ontwikkelen, maar beperkt tot een toetsing van de haalbaarheid ervan. De reikwijdte van het pilot codeboek is daarom bewust beperkt, om - gezien het arbeidsintensieve coderingsproces - deze stap in het onderzoek beheersbaar te houden. De selectie van indicatoren die in het pilot codeboek zijn opgenomen is daarom beperkt tot indicatoren die van belang zijn voor de kwalificatie van rechts-extremistische, jihadistische en impliciete extremistische content. Voor terroristische content gaf het team prioriteit aan typen content die het meest voor discussie vatbaar zijn, terwijl voor illegale content alleen vormen met een mogelijke overlap met impliciete extremistische content werden meegenomen.

Voor de kwalificatie van terroristische content richtte de pilot zich op drie soorten misdrijven: aanzetten tot het plegen of deelnemen aan een terroristisch misdrijf, verheerlijking van terroristische misdrijven en rekrutering voor een terroristische organisatie. De indicatoren zijn ontleend aan de EU-verordening terroristische online-inhoud (EU 2021/784), die bindend is voor zowel platforms als bevoegde autoriteiten binnen de EU. Elk bericht werd gescreend op basis van deze criteria om te beoordelen of het als terroristische content kon worden aangemerkt.

Voor de kwalificatie van illegale content werd slechts een beperkte set niet-terroristische categorieën opgenomen om de haalbaarheid van het duidingskader te testen: aanzetten tot haat, aanzetten tot geweld, en ontkenning, bagatellisering of vergoelijking van internationale misdrijven. Deze werden geselecteerd op basis van hun potentiële overlap met extremistische narratieven en beoordeeld met behulp van indicatoren uit het Nederlandse Wetboek van Strafrecht.

In beide gevallen zorgde een extra waarborg ervoor dat content die onder de vrijheid van meningsuiting valt niet per ongeluk als terroristisch of illegaal werd geclassificeerd. Voor deze laatste stap werd het belangrijk geacht te toetsen of er niet sprake was van satire, een parodie, een artistieke uiting, een gerechtvaardigde bijdrage aan het publieke debat, of een legitieme referentie naar een herdenking van een historisch feit, het koloniale verleden, of dekolonisatie.

Indicatoren van Impliciete Extremistische Content

Waar het onderzoeksteam voor de vorige categorieën kon verwijzen naar wettelijke kaders om essentiële indicatoren te identificeren, was dit niet het geval voor impliciete extremistische content. Het onderzoek heeft daarom, op basis van een analyse van literatuur en beleidsrapporten, verschillende relevante indicatoren voor het classificeren van impliciete extremistische content geïdentificeerd. De voorgestelde categorieën van indicatoren om impliciete extremistische content te identificeren zijn:

- **Verhulling van Betekenis (CM)**: opzettelijke verhulling via ironie, humor of codetaal.
- **Schadelijke Allianties/Affiliaties (HA)**: verwijzingen naar extremistische groepen, ideologieën of symbolen.
- **Problematische Verwijzingen (PR)**: het aanhalen van historische of actuele gebeurtenissen met een extremistische framing.
- **Impliciete Actietriggers (AT)**: subtiele signalen die het publiek aansporen om actie te ondernemen.
- **Verondersteld Oogmerk Schade te berokkenen (IH)**: verondersteld oogmerk schade te berokkenen met de geplaatste content.

Elke categorie bestond uit meerdere verschillende indicatoren. Om content te kwalificeren als impliciete extremistische content, moest er een combinatie van indicatoren aanwezig zijn. Ook hier werden aanvullende waarborgvragen ingebouwd, die toetsen of er sprake was van satire, een parodie, een artistieke iting, een gerechtvaardigde bijdrage aan het publieke debat, of een legitieme referentie naar een herdenking van een historisch feit, het koloniale verleden, of dekolonisatie, om er op deze wijze voor te zorgen dat content die onder de vrijheid van meningsuiting valt, werd uitgezonderd.

De indicatoren van terroristische, illegale en impliciete extremistische content werden opgenomen in een pilot codeboek om hun betrouwbaarheid te testen (onderzoeksvragen 3 en 4). De indicatoren werden beoordeeld op drie criteria: bruikbaarheid, nuttigheid en interpreteerbaarheid.

Bruikbaarheid (*usability*) weerspiegelt hoe eenvoudig en tijdsefficiënt het is om een indicator te coderen. Groen betekent gemakkelijk, oranje tot op zekere hoogte veeleisend en rood wordt gebruikt voor lastige indicatoren. *Nuttigheid* (*usefulness*) geeft aan of een indicator een betekenisvolle bijdrage levert aan de algehele beoordeling. Rood duidt op overbodigheid en oranje geeft aan dat er beperkte bijdrage of overlap is. *Interpreteerbaarheid* (*interpretability*) evalueert de mate van subjectiviteit van de codering. Groen staat voor duidelijke indicatoren, oranje voor indicatoren die tot op zekere hoogte subjectief zijn en verfijning vereisen, en rood voor zeer subjectieve indicatoren, geïdentificeerd door discrepanties tussen codeurs.

Het onderzoek reflecteert vervolgens op de resultaten van deze evaluatie die de volgende bevindingen opleverde:

Het Belang van Context

Indicatoren functioneren zelden geïsoleerd. Interpretatie hangt sterk af van de context, waaronder de identiteit van de gebruiker, culturele referenties en hoe de content door gebruikers wordt ervaren. Deze contextafhankelijkheid compliceert opschaling van detectiemechanismen en ondermijnt de betrouwbaarheid wanneer er uitsluitend gebruik wordt gemaakt van geautomatiseerde systemen.

Hybride Detectiemodellen

Het onderzoek onderstreept dat geautomatiseerde hulpmiddelen de nuance van impliciete extremistische content niet kunnen herkennen. Een hybride model - een combinatie van AI-ondersteunde voorselectie en menselijke expertise - is essentieel. Menselijke codeurs brengen contextgevoeligheid, maar hebben gestructureerde begeleiding, training en waarborgen tegen bias nodig.

Operationele Keuzes van de Technologiesector

Platforms vertrouwen in toenemende mate op geautomatiseerde hulpmiddelen als ruggengraat van hun moderatiesystemen en presenteren dit als oplossing voor schaal- en efficiëntieproblemen. Uit het onderzoek blijkt echter dat, hoewel automatisering nuttig is voor het detecteren van openlijk terroristische of illegale content, het slecht presteert bij toepassing op impliciete extremistische content dat vaak gebaseerd is op culturele nuances, ironie of gecodeerde referenties. De afname van menselijke moderators bij grote platforms verergert dit risico, waardoor er een groeiende kloof ontstaat tussen de complexiteit van schadelijke content en de methodes die de sector heeft gekozen om dit aan te pakken.

Transparantie en Verantwoording

Het gebrek aan transparantie van platforms over moderatiepraktijken beperkt het publieke vertrouwen en de academische controle ernstig. Er zijn diverse moderatiemogelijkheden voor zowel content als accounts, waaronder het beperken van de zichtbaarheid. Iedere optie heeft een andere invloed op de vrijheid van meningsuiting. Het blijft echter onduidelijk hoe vaak welke moderatiebeslissingen worden genomen en in welke mate rekening wordt gehouden met de impact die deze beslissingen hebben op de vrijheid van meningsuiting (proportionaliteitsvereiste). Zonder duidelijke rapportage en toegankelijke beroepsmogelijkheden blijven gebruikers in het ongewisse over de regels die hun online uitingen beheersen.

Haalbaarheid van een Duidingskader?

Op basis van de bevindingen zijn er verschillende fundamentele uitdagingen geïdentificeerd. Gezamenlijk wijzen deze uitdagingen erop dat, hoewel stapsgewijze verbeteringen mogelijk zijn, een universeel en volledig betrouwbaar duidingskader op dit moment onhaalbaar is.

Uitdagingen

Met betrekking tot de haalbaarheid van een betrouwbaar duidingskader werden de volgende uitdagingen geïdentificeerd:

- **Meerduidige Definities en Vage Grenzen**
  - Gebrek aan universeel aanvaarde wettelijke definities van termen als ‘terrorisme’, ‘aanzetten tot haat’, ‘gewelddadig extremisme’ of ‘opruiing’.
  - Gebrek aan internationaal erkende definities voor essentiële begrippen zoals ‘groep’, ‘legitiem’ of ‘zelfverdediging’.
  - Onduidelijke referentiekaders voor wanneer content als schadelijk versus aanvaardbaar wordt beschouwd.
  - Lastig om onderscheid te maken tussen extremistische retoriek en satire, politieke kritiek of legitiem debat.
  - Vage categorieën ondermijnen consistentie in codering en overeenkomst tussen codeurs.
  - Het impliciete karakter van content kan betrekking hebben op het verhullen van de schadelijkheid, rechtmatigheid of illegaliteit van de content. Begrijpen wanneer dit opzettelijk gebeurt is lastig.

- **Veranderende Verhullingstactieken**
  - Extremistische actoren passen taalgebruik en strategieën snel aan, waardoor statische duidingskader achterhaald raken.
  - *Dog whistles*, ironie, memes en gecodeerde termen ontlopen detectie.

- **Risico Subjectiviteit en Vooringenomenheid bij Menselijke Moderatoren**
  - De sterke afhankelijkheid op persoonlijke interpretatie en context leidt tot inconsistente resultaten.
  - Risico dat gewone uitingen van ongenoegen of oppositie abusievelijk als extremistisch worden geclassificeerd.
  - De culturele, ideologische of politieke achtergronden van codeurs kunnen hun oordelen beïnvloeden.
  - Indicatoren zijn kwetsbaar voor misbruik door bevooroordeelde codeurs of *flaggers*.

- **Nauwkeurigheid v. Vooringenomenheid**
  - Geautomatiseerde moderatie implementeert de vastgestelde algoritmes weliswaar nauwkeuriger, maar het risico bestaat dat er sprake is van ingebouwde vooringenomenheid die zich niet snel manifesteert.
  - Geautomatiseerde moderatie is weliswaar goedkoper en sneller, maar vaak minder goed in staat om impliciete extremistische content accuraat te detecteren, daarentegen is menselijke beoordeling beter geschikt voor de beoordeling van impliciete extremistische content, maar is het tevens duurder en brengt het risico van vooringenomenheid met zich mee.
  - Het enorme volume aan online content overstijgt de capaciteit van zowel menselijke codeurs als huidige AI-systemen.

- **Contextafhankelijkheid en Arbeidsintensiteit**
  - Veel indicatoren vereisen diepgaande kennis van gebruikersgeschiedenis, ideologie of platformdynamiek. Dit beperkt de bruikbaarheid en maakt identificatie lastig.

Een bijkomende structurele uitdaging zijn de onderliggende bedrijfsmodellen die de praktijken van hostingdiensten (HSP's) vormgeven. Grote technologiebedrijven zijn primair winstgedreven en hebben weinig behoefte om te investeren in menselijke moderatoren die de betrouwbaarheid zouden verbeteren en fundamentele rechten zouden kunnen waarborgen. Dit onderzoeksteam stuitte, net als vele anderen, op aanzienlijke terughoudendheid van platforms om mee te werken of transparantie te bieden over hun moderatieaanpak. Dit onderstreept de verantwoordingskloof tussen private zelfregulering en het publieke belang. De inkrimping van menselijke moderatieteams, gecombineerd met de ondoorzichtigheid van algoritmische besluitvorming, ondermijnt niet alleen de betrouwbaarheid van detectie, maar beperkt ook de mogelijkheden voor democratisch toezicht.

Aanbevelingen

Een universeel geldig en betrouwbaar duidingskader voor het detecteren van impliciete extremistische content lijkt niet haalbaar (onderzoeksvragen 5-7). Meer betrouwbare en adaptieve kaders zouden echter wel haalbaar kunnen zijn voor gebruik door hostingdiensten (HSP's). Door definities te verfijnen, complexe indicatoren te operationaliseren, en iteratief leren, hybride systemen en gezamenlijk toezicht in te bedden, kunnen de gebruikte indicatoren worden omgevormd tot consistentere, nuttigere en betrouwbaardere hulpmiddelen. Een dergelijke aanpak balanceert de noodzaak om impliciete extremistische content te identificeren die de vrijheid van meningsuiting waarborgt. Uiteindelijk, en zolang er geen aanvullende wettelijke kaders van toepassing zijn, zal de effectiviteit van elk detectiemechanisme afhangen van de bereidheid en capaciteit van HSP's om het verantwoord toe te passen. Overheden zouden, in samenwerking met de EU, hun dialoog met HSP's moeten intensiveren om dit proces te stimuleren.

Of een dergelijk universeel duidingskader überhaupt wenselijk is, blijft een open vraag. Ethische overwegingen moeten een centrale rol spelen bij die afweging. Daarbij speelt het feit dat onze huidige vorm van communicatie steeds complexer is geworden: online- en offline-werelden zijn nauw verbonden, en zijn een weerspiegeling van zowel sociale diversiteit als groeiende polarisatie. Deze snelle transformatie heeft het publieke debat over welke normen en waarden gelden op sociale mediaplatforms ingehaald. Gezien het feit dat wettelijke kaders ambigu of inconsistent zijn in het stellen van grenzen, zou de ontwikkeling van ieder duidingskader voor online content moeten beginnen met een brede maatschappelijke dialoog over wat acceptabele uitingsvormen zijn en wanneer dit overgaat naar onacceptabel en schadelijk uitingsvormen.

Ondanks deze bezwaren benadrukken we hieronder aanbevelingen voor de technologiesector en voor beleidsmakers. Deze aanbevelingen volgen voor het merendeel rechtstreeks uit de bevindingen van de studie. Echter, de expertise en praktische ervaring die de onderzoekers hebben opgedaan bij het implementeren van preventieprogramma's en bij capaciteitsopbouw, hebben ook bijgedragen aan de formulering van enkele aanbevelingen.

Basisvoorwaarden die HSP's kunnen implementeren voor een betrouwbaar en accuraat duidingskader:

1. **Duidelijke Definities en Referentiekaders**
   a. Leef de verplichting na om precieze definities voor vage termen te hanteren;
   b. Stel referentiekaders vast voor opruiing, haat of vijandigheid, die niet in strijd zijn met de vrijheid van meningsuiting;
   c. Wees transparant over de combinatie van indicatoren die aanwezig moeten zijn voor de vaststelling van impliciete extremistische content.

2. **Richtlijnen en Voorbeelden**
   a. Geef illustratieve voorbeelden van wat wel en niet geoorloofde content is ten aanzien van het gehele ideologische spectrum, inclusief satire, geoorloofde kritiek en schadelijke retoriek.
   b. Gebruik beslisbomen of coderingsstroomdiagrammen om toepassing te standaardiseren.
   c. Ontwikkel typologieën van wat of wie tot de *in- or out-group* behoort om accurate toepassing van de betreffende indicator te vergemakkelijken.

3. **Beperking van Subjectiviteit en Vooringenomenheid**
   a. Voorzie codeurs van training, prompts en strategieën om vooringenomenheid tegen te gaan.
   b. Pas beoordelingsprotocollen met het "vierogenprincipe" en consensusmethoden toe voor de categorie content die gekwalificeerd kan worden als impliciete extremistische content.

4. **Operationalisering van Complexe Indicatoren**
   a. Splits brede indicatoren op per ideologische stroming en in subcategorieën.
   b. Gebruik triangulatie met contextuele aanwijzingen (bijv. geschiedenis, platformdynamiek) om betrouwbaarheid te verbeteren.

5. **Iteratieve en Adaptieve Kaders**
   a. Gebruik voor de verschillende ideologische stromingen aparte duidingskaders, in plaats van een universeel duidingskader.
   b. Beschouw duidingskaders als "levende documenten" die mee ontwikkelen met nieuwe extremistische tactieken.
   c. Update de duidingskaders regelmatig aan de hand van nieuwe inzichten en voorbeelden uit onderzoek of naar aanleiding van resultaten die uit evaluaties komen.

6. **Hybride AI-Menselijke Systemen**
   a. Combineer AI-voorselectie voor schaalbaarheid met deskundige menselijke beoordeling voor contextgevoelige beoordelingen.
   b. Gebruik LLM's en databanken om arbeidsintensieve taken uit te voeren, maar houd mensen ondertussen betrokken bij het proces.

7. **Multistakeholder-Samenwerking**
   a. Werk regelmatig samen met onderzoekers, praktijkdeskundigen, platforms, het

maatschappelijk middenveld en gemarginaliseerde groepen bij de ontwikkeling en herziening van duidingskaders.
   b. Bouw een brede consensus om bias, overschrijding en misclassificatie te beperken.

8. **Transparantie en Beroepsprocedures**
   a. Verbeter transparantie en rapportage over moderatiebeslissingen.
   b. Respecteer het proportionaliteitsbeginsel bij moderatiebeslissingen om de vrijheid van meningsuiting beter te waarborgen: Om de vrijheid van meningsuiting te waarborgen, is het nodig om een breder scala aan proportionele moderatieopties te ontwikkelen en ook daadwerkelijk toe te passen in de praktijk.
   c. Geef heldere informatie over hoe en waar men in beroep kan gaan tegen een moderatiebesluit.

Specifieke Aanbevelingen voor de Beleidsmakers

Dit onderzoek is uitgevoerd op verzoek van de NCTV. Hoewel de bevindingen en aanbevelingen van dit onderzoek relevant zijn voor een breder publiek, speelt de NCTV, als coördinerende actor, een sleutelrol in het vormgeven en implementeren van beleid in Nederland, in de samenwerking met Europese partners om het Europese beleid te bevorderen en in de dialoog met HSP's. Op basis van de bevindingen van dit onderzoek en gebaseerd op de expertise van de leden van het onderzoeksteam, formuleren wij daarom aanbevelingen voor beleidsmakers die in het bijzonder een coördinerende rol vervullen, zoals de NCTV en de ATKM.

Beleidsmakers wordt aanbevolen om:

1. **Af te zien van de term 'borderline' content**, omdat dit de verwarring over de reikwijdte en betekenis van de term alleen maar zal vergroten.

2. **Een *multistakeholdergroep* op te richten**, bestaande uit onderzoekers, praktijkexperts, platforms, het maatschappelijk middenveld en gemarginaliseerde groepen om (regelmatig) en op een transparante wijze te reflecteren op en publiekelijk te rapporteren over:
   a. Een reeks indicatoren om terroristische, illegale en impliciete extremistische content te detecteren, in lijn met het idee van het bijhouden van een levend document;
   b. De indicatoren die volgens dit onderzoek als problematische worden beschouwd en hun formulering te verbeteren;
   c. De kritische grens voor de combinatie van indicatoren die nodig zijn voor de kwalificatie als impliciete extremistische content
   d. Kenmerkende gebeurtenissen of historische feiten, specifiek voor de Nederlandse context, alsmede veelgebruikte uitdrukkingen, specifiek taalgebruik, en codetaal dat eigen is aan de Nederlandse taal en gehanteerd wordt door extremistische groepen die online content in het Nederlands plaatsen, wat van belang kan zijn voor de contextuele interpretatie van impliciete extremistische content.

3. **Een publiek debat te bevorderen**
   a. Over wat schadelijke en onrechtmatige content is en wat schadelijke maar rechtmatige content is;
   b. Over hoeveel autonomie HSP's hebben bij het beheren van onze openbare ruimte voor publieke meningsuiting.

4. **Media-geletterdheid programma's op scholen, preventieprogramma's voor jongerenorganisaties en strategische communicatieve de-escalatietechnieken te ondersteunen**, die inzetten op het tegengaan van verspreiding van dergelijke schadelijke content, en het vergroten van weerbaarheid, in het bijzonder voor minderheids- en gemarginaliseerde groepen.

5. **Een helder strategisch communicatiebeleid te ontwikkelen** om adequaat te reageren bij grootschalige verspreiding van schadelijke maar rechtmatige content, inclusief uitleg waarom iets als schadelijk wordt beschouwd. Tevens **publiekelijk stelling te nemen tegen schadelijke onrechtmatige content die gericht is op gemarginaliseerde groepen.** Daarnaast is het van belang lokale beleidsmakers hierin te ondersteunen.

De coördinerende overheidsactoren (zoals de NCTV of de ATKM) worden aanbevolen om in dialoog met HSP's

6. **Op transparante wijze samen te (blijven) werken met grote en kleine hostingdiensten:**
   a. Om een open discussie te voeren over de indicatoren die zij gebruiken in hun duidingskaders en of zij verschillende duidingskaders gebruiken voor verschillende ideologieën;
   b. Om informatie-uitwisseling en transparantie over de toepassing van verschillende proportionele moderatiebeslissingen en de impact op de vrijheid van meningsuiting te bevorderen.

7. Zonder de HSP's van hun primaire verantwoordelijkheid te ontdoen, maar rekening houdend met het feit dat het Nederlands een kleine taal is**, een lijst met veelgebruikte uitdrukkingen, specifiek taalgebruik, codetaal die specifiek zijn voor het Nederlands** en worden gebruikt door extremistische groeperingen die online actief zijn in Nederland - zoals besproken in de multistakeholdergroep - **te delen,** om zo te helpen bij de contextuele interpretatie van impliciete extremistische content.

8. **Regelmatig contextuele achtergrondinformatie te verstrekken** aan HSP's over typisch Nederlandse (actuele of historische) gebeurtenissen, die hen kunnen helpen bij de contextuele interpretatie van online content.

De coördinerende overheidsactoren wordt aanbevolen, om in dialoog met andere Europese lidstaten en de Europese Commissie:

9. **Samen met de sector een sectorbrede gedragscode te ontwikkelen,** die een keurmerk aanbiedt dat consumenten een beter inzicht biedt in hoe goed HSP's scoren op detectie en moderatie; die normen vaststelt voor het percentage van menselijke beoordelingen en moderatiebeslissingen; en die duidelijkheid schept over de termen die in de ToS gebruikt worden, en ten aanzien van de transparantie van moderatiebeslissingen en de beroepsprocedures.

10. **Zich in te zetten voor verdere versterking van Europese regelgevingskaders** door meer transparantie en verantwoordingsplicht van HSP's te eisen, een ex-ante evaluatie te eisen over hoe HSP's de vrijheid van meningsuiting respecteren door hun ToS toe te passen, regelmatige ex-post evaluaties te eisen over hoe die vrijheid is gerespecteerd in moderatiebeslissingen, strengere regels te hanteren voor moderatiemethodes (AI versus handmatig), en door heldere definities en richtlijnen te bieden.

Conclusie

Het onderzoek concludeert dat de ontwikkeling van een universeel, betrouwbaar duidingskader voor de detectie van terroristische, illegale en impliciete extremistische online content niet haalbaar is gezien de huidige juridische, technische en ethische omstandigheden. Meerduidigheid van definities, contextuele complexiteit en de steeds veranderende tactieken van extremistische actoren ondermijnen de betrouwbaarheid van dergelijke kaders en vergroten de risico's voor de vrijheid van meningsuiting.

Toch betekent de onhaalbaarheid van een universeel duidingskaders niet dat vooruitgang onmogelijk is. Stapsgewijze verbeteringen zijn mogelijk door duidelijkere definities, adaptieve indicatoren, hybride AI-menselijke modellen en samenwerking tussen *multistakeholders'*. In plaats van te zoeken naar een pasklare oplossing, zouden platforms en beleidsmakers flexibele, dynamische aanpakken moeten nastreven die afgestemd zijn op specifieke contexten en ideologieën.

Voor de NCTV en de bredere beleidswereld reikt de uitdaging verder dan alleen detectie, maar raakt het ook aan bredere maatschappelijke vraagstukken. Wat wordt beschouwd als schadelijke maar rechtmatige (*awful but lawful*) content? Hoeveel autonomie mogen private bedrijven hebben in het besturen van online publieke ruimtes? Hoe kunnen samenlevingen een balans vinden tussen veiligheid en vrijheid?

De weg voorwaarts vereist voortdurende dialoog, transparantie en adaptiviteit. Naarmate extremistische actoren blijven innoveren, moeten beleidsmakers, platforms en gemeenschappen dat ook doen. Weerbaarheid opbouwen vereist niet alleen technologische oplossingen, maar ook een democratisch debat over de normen van onlinecommunicatie en de waarden die samenlevingen willen handhaven.

International Centre for Counter-Terrorism

International Centre for Counter-Terrorism (ICCT)
T: +31 (0)70 763 0050
E: m.e.centre@icct.nl
https://icct.nl/Monitoring-and-Evaluation