Online extremisme en radicalisering: Verkenning van nieuwe detectie- en interventiemogelijkheden (P102)

ONGERUBRICEERD Releasable to the public ⟩ TNO 2024 R10310
Maart 2024

Defence, Safety & Security
www.tno.nl
+31 88 866 10 00
info@tno.nl

TNO 2024 R10310 – Maart 2024

Online extremisme en radicalisering: Verkenning van nieuwe detectie- en interventiemogelijkheden (P102)

Auteurs	E. Poot, D. Blok, A. de Vries, S. Raaijmakers, D. Van Hemert
Rubricering rapport	ONGERUBRICEERD Releasable to the public
Vastgesteld door	NCTV
Vastgesteld d.d.	18 maart 2024
Titel	ONGERUBRICEERD Releasable to the public
Managementuittreksel	ONGERUBRICEERD Releasable to the public
Samenvatting	ONGERUBRICEERD Releasable to the public
Rapporttekst	ONGERUBRICEERD Releasable to the public
Bijlagen	ONGERUBRICEERD Releasable to the public
Aantal pagina's	73 (excl. voor- en achterblad en distributielijst)
Aantal bijlagen	2
Opdrachtgever	Ministerie van Justitie en Veiligheid, NCTV
Programmanaam	NCTV-KOP
Programmanummer	P102
Projectnaam	Aanpak online extremisme
Projectnummer	060.55233

Managementuittreksel

Online extremisme en radicalisering: Verkenning van nieuwe detectie- en interventiemogelijkheden

Programma

Programmanaam:
NCTV-KOP

Programmanummer: P102

Programmaplanning: Startdatum 1 januari 2023
Einddatum 1 april 2024

Programmabegeleider:
Ministerie van Justitie en Veiligheid, NCTV

Programmaleider:
Dr. M.P.W. van Berlo
TNO Defence, Safety & Security

Project

Projectnaam:
Aanpak online extremisme

Projectnummer: 060.55233

Projectplanning: Startdatum 13 juli 2023
Einddatum 15 maart 2024

Projectbegeleider:
Ministerie van Justitie en Veiligheid, NCTV

Projectleider:
Drs. J. de Schepper
TNO Defence, Safety & Security

Probleemstelling

In dit onderzoek richt TNO zich op mogelijkheden voor het detecteren en interveniëren op online extremisme en radicalisering, in lijn met de Nationale Contraterrorisme Strategie 2022-2026, die is gericht op het voorkomen en aanpakken van terrorisme en gewelddadig extremisme. De onderzoeksvraag in dit project luidde: Welke innovatieve technologische oplossingen zijn effectief voor a) de signalering/detectie en b) het tegengaan van de verspreiding van radicale en (rechts)extremistische content online en onder welke condities kunnen deze worden toegepast?

Beschrijving van de werkzaamheden

Wij hebben literatuuronderzoek uitgevoerd naar detectie en tegengaan van online rechtsextremisme, interviews gevoerd met zowel publieke als private partijen, zoals overheidsinstellingen, technologiebedrijven, sociale mediaplatformen of hosting providers, en een serie workshops met deze stakeholders gehouden. Op basis van de resultaten komen wij tot een overzicht van huidig gebruikte en state-of-the-art innovatieve technologische oplossingen ten behoeve van detectie, monitoring en interventies om de verspreiding van online extremistische content tegen te gaan. Onze focus hierbij was rechtsextremisme.

Resultaten en conclusies

Methoden die mogelijk toegepast kunnen worden voor detectie van extremisme zijn het detecteren van taalgebruik en psychologische en gedragssignalen,sociale netwerkanalyse (SNA) en de inzet van technologie als machine learning (ML) of deep learning (DL). Interessante nieuwe ontwikkelingen voor detectie van extremistische inhoud zijn Large Language Models (LLM's) en AI-gebaseerde detectie en interpretatie van beelden. Deze technologieën kunnen daarnaast ook ingezet worden voor het maken van extremistische content.
Ondanks veelbelovende technologische ontwikkelingen speelt de mens nog altijd een zeer belangrijke rol in de detectie en interpretatie van online extremistische content. Vooral bij overheidspartijen wordt veel informatie min of meer handmatig gezocht, en er wordt nog relatief weinig gewerkt met artificiële intelligentie (AI). De opvolging van door AI gedetecteerde extremistische patronen in data is veelal afhankelijk van menselijke besluitvorming, behalve in overduidelijke gevallen, zoals het gebruik van hashfuncties om

unieke content te herkennen die eerder als extremistisch of terroristisch werd aangeduid en verwijderd; slechts dan wordt door private partijen soms tot automatische verwijdering overgegaan. De nieuwe AI-gebaseerde mogelijkheden voor de detectie en interpretatie van online extrremistische context enerzijds en het zwaarwegende belang van menselijke interpretatie van en controle op algoritmen anderzijds roepen om nieuwe en zorgvuldig getoetste werkverdelingen tussen mens en algoritme. De rol van AI bij online interventies is nog beperkt, en verdient aanvullend empirisch onderzoek: op welke manier kan AI bijdragen aan meetbaar effectieve interventies?

Toepasbaarheid

Interventies kunnen zowel proactief als reactief zijn, en zowel offline als online worden ingezet. Voorbeelden van proactieve interventies zijn het investeren in mediawijsheid, counterspeech of het markeren van content als waarschuwing richting eindgebruikers dat deze mogelijk onwaar of radicaliserend of extremistisch is. Reactieve interventies gaan vooral over online contentmoderatie of casusgericht fysiek ingrijpen door de overheid. Een nieuwere reactieve interventie bestaat uit het selectief uitzetten van functies in een online platform zodanig dat radicale en extremistische groepen gehinderd worden om hun doelen te bereiken, zoals financiering of rekrutering. Naast gerichte interventies zijn er steeds meer interventies die als een breder barrièremiddel gezien kunnen worden, zoals ingrijpen op desinformatie en misinformatie, achterliggende verdienmodellen en cross-platformgebruik, waarbij op diverse online platforms aan elkaar worden gekoppeld voor effectieve verspreiding van extremistische content.

In termen van technologie zijn er nieuwe kansen op het gebied van de inzetbaarheid van bijvoorbeeld Large Language Modellen (LLMs) en de hierop gebaseerde AI-assistenten als detectoren van radicale of extremistische content. Deze modellen en assistenten kunnen analyses ondersteunen en daarbij een mens-leesbare uitleg produceren. Ook kunnen ze via natuurlijke taal bevraagd worden door analisten. Verder is automatische contentmoderatie met LLMs mogelijk, zoals het herformuleren van radicale of extremistische teksten tot neutrale teksten, of het formuleren van counter-narratives die een tegengeluid laten horen. Systematisch onderzoek naar gebruik en kwaliteit van deze modellen in de context van online extremisme en radicalisering ontbreekt vooralsnog.

Stakeholders zijn zich slechts in beperkte mate bewust van nieuwe AI-gebaseerde dreigingsbeelden rond online radicalisering en extremisme, en zijn beperkt in staat hierop te handelen, zowel voor wat betreft detectie als opvolging. Een belangrijke reden voor beperkte opvolgingsmogelijkheden is wetgeving die achterloopt op de snelle technologische ontwikkelingen. Er is dan ook behoefte aan uitzonderingsclausules om effectiever te kunnen opereren. Daarnaast willen zowel publieke als private partijen van voornamelijk reactief interveniëren naar een meer proactieve interventiestrategie. Een meer integrale aanpak waarbij ook steeds meer sociale partners, die zich op preventie richten, een rol spelen is wenselijk. Meer afstemming (of synchronisatie) in maatregelen om tot een meer integrale aanpak te komen is daarbij noodzakelijk. Er zijn recentelijk nieuwe organisaties en samenwerkingsverbanden ontstaan die een specifiekere rol pakken in het publiek-private speelveld. De Autoriteit Online Terroristisch en Kinderpornografisch Materiaal (ATKM) in Nederland, het Global Internet Forum to Counter Terrorism (GIFCT) en Tech Against Terrorism zijn daar mooie voorbeelden van. Voorbeelden van taken die deze partijen op zich nemen zijn de coördinatie van interventies, verbetering van onderlinge interventieprocessen, kennisdeling, het op weg helpen van (nieuwe) platformen en het (samen) ontwikkelen van detectietechnologie. Wet- en regelgeving maar ook imago of legitimiteit en onderling

vertrouwen spelen een belangrijke rol bij de samenwerking tussen de verschillende actoren die actief zijn in het bestrijden van online extremisme en radicalisering. Op diverse niveaus, van operationeel tot tactisch en strategisch zal er meer kennis, inzicht, ervaringen en data uitgewisseld moeten worden , met name voor wat betreft de effectiviteit van bepaalde interventies op online extremistische content en radicalisering.

Samenvatting

De minister van Justitie en Veiligheid heeft aan de Tweede Kamer toegezegd om medio 2023 te komen met een nadere strategie voor de aanpak van radicalisering en extremisme, met daarbij bijzondere aandacht voor preventie, mede als aanvulling op de implementatie van de EU Verordening 2021/784 inzake het tegengaan van terroristische online inhoud. Dit onderzoek beoogt hieraan bij te dragen door met name het handelingsperspectief van de overheid rond online radicalisering en extremisme te verkennen. Hiertoe biedt het onderzoek een overzicht van huidig gebruikte en state-of-the-art innovatieve technologische oplossingen voor detectie, monitoring en interventies om de verspreiding van online extremistische content tegen te gaan. Daarnaast geeft het onderzoek technische en strategische beleidsaanbevelingen.

De onderzoeksvraag in dit project luidt:

Welke innovatieve technologische oplossingen zijn effectief voor a) de signalering/detectie en b) het tegengaan van de verspreiding van radicale en (rechts)extremistische content online en onder welke condities kunnen deze worden toegepast?

Aanpak
Voor het beantwoorden van de onderzoeksvraag is een literatuuronderzoek uitgevoerd en zijn interviews en workshopsessies gehouden met publieke en private partijen. Het literatuuronderzoek beoogde een state-of-the-art overzicht te geven van detectie- en interventiemogelijkheden op het gebied van online rechtsextremistische en radicale content. De interviews en workshopsessies met publieke als private partijen, zoals overheidsinstellingen, technologiebedrijven, sociale mediaplatformen en hosting providers, en NGO's die onafhankelijk kennis delen in publieke, private, of publiek-private samenwerkingsverbanden, dienden om een beeld te krijgen van de huidige praktijk. De behoeften van alle stakeholders zijn geïnventariseerd en gekoppeld aan nieuwe mogelijkheden zoals die uit de literatuur en vanuit de partijen naar voren kwamen. Verder zijn de geïdentificeerde behoeften en mogelijkheden door de betrokken publieke en private partijen geprioriteerd voor de korte en lange termijn.

Detectiemogelijkheden uit de literatuur en de praktijk
Methodieken uit de literatuur die mogelijk toegepast kunnen worden voor detectie van extremisme zijn onder andere gebaseerd op taalgebruik, psychologische signalen en gedragssignalen, maar ook sociale netwerkanalyse (SNA) en de inzet van technologie als machine learning (ML) of deep learning (DL) zijn relevant. Daarnaast vormen Large Language Models (LLMs) een recente technologie waarmee inhoud van extremistische en radicale teksten en beelden beter en meer uitlegbaar gedetecteerd zouden kunnen worden. Deze technologie is echter ook 'dual use', in de zin dat het kan worden ingezet voor het maken van extremistische content, net als verwante "generatieve AI" modellen voor beeldgeneratie.
Uit de interviews en workshopsessies komt een beeld van de huidige detectiepraktijk naar voren waarin de mens nog altijd een zeer belangrijke rol. Overheidspartijen gebruiken relatief eenvoudige technologieën om informatie te zoeken in grote hoeveelheden data (zoals bijvoorbeeld door de politie in een casusgerichte aanpak). Deze partijen werken over het algemeen nog relatief weinig met artificiële intelligentie (AI) voor de detectie van extremistische en radicale content. Private sociale mediaplatformen werken weliswaar veel

meer met AI, bijvoorbeeld voor het herkennen van inhoud (in tekst of beeld), of de analyse van patronen in (meta)data, maar vrijwel altijd speelt de mens een cruciale rol in het *beoordelen of verwijderen* van content of accounts. Alleen in overduidelijke gevallen, zoals het herkennen van unieke content via “hashes” (unieke digitale vingerafdrukken) van content die eerder als extremistisch of terroristisch werd geïdentificeerd wordt tot automatische verwijdering overgegaan. Bij een beperkt aantal platformen dat verder gaat in het automatisch verwijderen kunnen eindgebruikers altijd bezwaar maken tegen verwijdering en zijn de contentmoderatieteams continu bezig om de detectiealgoritmen te verbeteren, mede op basis van de feedback van eindgebruikers..

Interventiemogelijkheden uit de literatuur en de praktijk
In zowel de literatuur als praktijk wordt onderscheid gemaakt tussen proactieve en reactieve interventies. Proactieve interventies zijn bijvoorbeeld het investeren in mediawijsheid, counterspeech of het markeren van content als mogelijk onwaar of extremistisch, zodat eindgebruikers gewaarschuwd worden. Reactieve interventies gaan vooral over online contentmoderatie (als uitkomst van detectie of meldingen) door platformen of fysiek ingrijpen door de overheid. Dit laatste betreft met name casusgerichte interventies door bijvoorbeeld de politie. Een nieuwere reactieve interventie is “digitaal hinderen”, zoals het selectief uitzetten van functies in een platform om radicale en extremistische groepen te belemmeren bij het bereiken van hun doelen, zoals financiering of cross-platform rekrutering.
Naast gerichte interventies zijn er steeds meer interventies die als een breder barrièremiddel gezien kunnen worden, zoals ingrijpen op des- en misinformatie.

Behoeften op basis van de literatuur en praktijk
Met betrekking tot detectie geven stakeholders aan zich slechts in beperkte mate bewust te zijn van nieuwe AI-gebaseerde dreigingsbeelden rond online radicalisering en extremisme, en beperkt in staat te zijn deze dreigingsbeelden op te volgen. Zowel overheidspartijen als private partijen geven aan de ontwikkelingen nauwgezet te willen volgen, maar het niet eenvoudig te vinden om gerichte actie te ondernemen op dit onderwerp, ook omdat SELP (Societal, Ethical, Legal, Privacy) aspecten complex zijn en (nieuwe) wetgeving zoals de AI Act voor Europa hier nu nog geen duidelijkheid biedt. Alle stakeholders vinden het wenselijk dat er meer transparantie komt over de uitlegbaarheid van detectiealgoritmen, vanwege integriteit richting samenleving en klanten en omdat uitlegbaarheid en bewijsvoeringmeer mogelijkheden biedt voor interventies en juridische vervolging Het is duidelijk dat wetgeving achterloopt op de snelle technologische ontwikkelingen en dat er behoefte is aan uitzonderingsclausules om effectiever te kunnen opereren, ondersteund door nieuwe samenwerkingsstructuren tussen publieke en private alle stakeholders.
Met betrekking tot interventies geven overheidspartijen aan behoefte te hebben aan digitale interventies, zowel voor zichzelf als in samenwerking met andere stakeholders. Alle partijen willen van een voornamelijk reactieve interventiestrategie naar een meer proactieve interventiestrategie. Hierbij is het wenselijk om meer samen te werken met sociale partners zodat op een integrale manier interventies kunnen worden ingezet.

Uitdagingen in detectie op basis van literatuur en praktijk
De opkomst van generatieve AI, die zelf content kan genereren, zoals tekst, geluid en stilstaand en bewegend beeld, biedt zowel nieuwe kansen als bedreigingen. Kansen zijn er op het gebied van de inzetbaarheid van bijvoorbeeld LLMs als detectoren van radicale of extremistische content. Deze modellen kunnen menselijke analyses ondersteunen doordat zij uitleg produceren en met natuurlijke taal bevraagd kunnen worden door analisten. Daarnaast liggen er kansen in automatische contentmoderatie met LLMs, zoals het herformuleren van radicale of extremistische teksten naar neutrale teksten of zelfs door het

formuleren van tegengeluiden (“*counter-narratives”)*. Systematisch onderzoek naar gebruik en kwaliteit van deze modellen in de context van online extremisme en radicalisering ontbreekt echter vooralsnog. De grote techpartijen spelen een actieve rol spelen in het ontwikkelen van deze taalmodellen en hebben daarom de meeste mogelijkheden om dit soort onderzoek uit te voeren. Desalniettemin kunnen ook op kleinere schaal, bijvoorbeeld met de nieuwe GPT-store van OpenAI[1] of met open source modellen, kleine, gespecialiseerde AI-assistenten getraind worden voor veilig gebruik binnen een organisatie. Generatieve AI vormt tegelijkertijd een bedreiging omdat het op grote schaal radicale of extremistische content kan produceren, volledig geautomatiseerd en op maat gemaakt voor bepaalde groeperingen (qua taal, stijl of onderwerp). Hierdoor komen geautomatiseerde detectie en interventie onder druk te staan. Zogenaamde *adversarial attacks* kunnen zelfs bestaande detectiemodellen ondermijnen omdat zij op basis van voorkennis van deze detectiemodellen content effectief kunnen versluieren. Tenslotte zijn er steeds meer platformen, zoals *Pastebin*, *Discord* of *Snapchat,* waarop in real-time informatie gedeeld wordt die slechts tijdelijk beschikbaar is (vluchtige, of *efemerische* content). Het bestaan van dergelijke platformen motiveert de ontwikkeling van nieuwe, *real time* detectietechnieken en interventiemethodieken.

Uitdagingen in interventies op basis van literatuur en praktijk

De publiek-private samenwerking rond online en offline interventies op het gebied van extremistische content staat nog in de kinderschoenen. De focus ligt voor de overheid meer bij het verkennen van online interventies in samenwerkingsconstructies, terwijl de online platformen en service providers zoeken naar interventies gericht op specifieke doelgroepen of problematiek. Daarnaast willen de platformen meer doen aan preventie van schadelijke content en van verspreiding te doen dan alleen aan contentmoderatie.

Zowel publieke als private partners vinden het wenselijk dat er, naast de meer reactieve aanpak van veiligheidspartners, een meer integrale aanpak komt waarin op preventie gerichte sociale partners betrokken zijn. Er zijn recentelijk nieuwe organisaties en samenwerkingsverbanden ontstaan die een specifiekere rol opnemen in het publiek-private speelveld rond online extremisme en radicalisering. De ATKM in Nederland, GIFCT en Tech Against Terrorism internationaal coördineren interventies, verbeteren onderlinge processen, delen kennis, helpen (nieuwe) online platformen op weg en ontwikkelen samen technologie. Daarnaast ontstaan nieuwe vormen van samenwerking vanuit private en publieke partijen zelf. Of dit soort samenwerkingen haalbaar en succesvol is, is voor een groot deel afhankelijk van wet- en regelgeving, imago, legitimiteit en onderling vertrouwen. Het gebrek aan onderzoek naar de effectiviteit van online interventies op extremistische en radicale content noopt in ieder geval tot meer uitwisseling van ervaringen en inzichten.

[1] https://openai.com/blog/introducing-gpts

Summary

The minister of Justice and Safety has promised the Parliament of the Second Chamber to come up with a further strategy for the approach of radicalization and extremism mid 2023, with special attention for prevention as an addition on the implementation of the EU Ordinance 2021/784 concerning the countering of terroristic online content. This research aims to contribute to this mainly by exploring the perspective of action from the government concerning online radicalization and extremism. For this purpose the research offers an overview of current used and state-of-the-art innovative technological solutions for detection, monitoring and interventions to counter the spread of online extremist content. Furthermore the research provides technical and strategic policy recommendations.

The research question in this project is:

Which innovative technological solutions are effective for a) the signaling/detection and b) the countering of the spreading of radical and (right-wing) extremist online content and under which conditions can these be applied?

Approach
In order to answer the research question a literature research has been carried out and interviews and workshop sessions have been organized with public and private parties. The literature research aimed at providing a state-of-the-art overview of detection and intervention possibilities on the domain of online right-wing extremist and radical content. The interviews and workshop sessions with public and private parties, like governmental institutions, technological companies, social media platforms, hosting providers, NGO's which share independently knowledge in public, private or public-private partnerships, served to get a picture of the current practice.
The needs of all stakeholders have been inventorized and linked to new possibilities as appeared from the literature and from the participating organizations. Furthermore the identified needs and possibilities have been prioritized by the concerned public and private organizations for the short and long term.

Detection possibilities from the literature and practice
Methodologies from the literature that might possibly be applied for detection of extremism are among things based on language use, psychological signals and behavioral signals, but also social network analyses (SNA) and the use of technology like machine learning (ML) or deep learning (DL) are relevant. In addition to that Large Language Models (LLM's) form a recent technology with which content of extremist and radical texts and images might be detected in an explainable way better and more frequently. This technology is however also 'dual use', in the sense that it can be applied for the production of extremist content, just as related 'generative AI' models for image generation. From the interviews and workshop sessions the image arises of the current detection practice in which the human still plays an important role. Governmental organizations use relatively simple technologies to search for information in large quantities of data (like for example by the police in case-oriented approaches). These organizations work in general still relatively few with artificial intelligence (AI) for detecting extremist and radical content. Even though private social media platforms work more with AI, for example in recognizing the content (text or image) or the analysis of patterns in (meta)data, the human almost always plays a crucial role in the judging or removing of content of accounts. Only in obvious cases, like the recognizing of

unique content via “hashes” (unique digital finger prints) of content which has previously been identified as extremist or terroristic, the automatic removal is proceeded. At a limited number of platforms which go further in the automatic removal, end users can always object to the removal. And the content moderation teams are continuously busy to improve the detection algorithms, partly on the basis of feedback from end users.

Intervention possibilities from the literature and practice
In literature as well as practice a distinction is made between proactive and reactive interventions. Proactive interventions are for example the investment in media literacy, counter speech or the marking of content as possibly fake or extremist, in order to warn end users. Reactive interventions are mainly about online content moderation (as a result of detection of notifications) by platforms or physical intervention by the government. The latter mainly concerns case oriented interventions by for example the police. A newer reactive intervention is ‘digital obstructing’, like the selectively switching off functions in a platform to obstruct radical and extremist groups in reaching their goals, like financing or cross-platform recruiting.
Next to focused interventions there are increasingly more interventions which can be seen as a broader means of barrier, like intervening on disinformation and misinformation.

Needs on the basis of literature and practice
With regard to detection, stakeholders mention that they are only limitedly aware of new AI based threat assessments regarding online radicalization and extremism, and limitedly capably to act on these threat assessments. Governmental organizations as well as private organizations mention that they are following the developments closely, but that it is not easy to take targeted action on this topic, also because SELP (Societal, Ethical, Legal, Privacy) aspects are complex and (new) legislation like the AI Act for Europe offers no clarity at this moment. All stakeholders consider it desirable that there will be more transparency about the explainability of detection algorithms, because of integrity towards society and customers and because explainability and evidence offer possibilities for interventions and legal persecution. It is clear that legislation is lagging behind the fast technological developments and that there is a need for exception clauses to be able to act more effectively, supported by new cooperation structures between public and private stakeholders.
With regard to interventions governmental organizations indicate the need for digital interventions, for themselves as well as in cooperation with other stakeholders. All the organizations want to go from a predominantly reactive intervention strategy towards a more proactive intervention strategy. Hereby it is desirable to cooperate more with social partners in order to deploy interventions in an integral manner.

Challenges in detection on the basis of literature and practice
The turn-out of generative AI, which can generate content on its own, like text, sound and stationary and moving images, offers new opportunities as well as threats. Opportunities exist on the domain of operationability of for example LLMs as detectors of radical or extremist content. These models can support human made analyses because they produce explanation and can be questioned with natural language by analysts. In addition opportunities exist in automatic content moderation with LLMs, like the reformulating of radical or extremist texts into neutral texts or even by formulating counter narratives. Systematic research in the use and quality of these models in the context of online extremism and radicalization is for now still nonexistent. The bigger tech organizations play an active role in the development of these language models and therefor have the most possibilities to carry out this kind of research. Nevertheless also on a smaller scale, for

example with the new GPT-store of OpenAI[2] or with open source models, small and specialized AI-assistants can be trained for safe use within an organization.
Generative AI forms at the same time a threat because of its ability to produce radical or extremist content on a large scale, fully automated and customized for certain groups (qua language, style or topic). Due to this automatic detection and intervention come under pressure. So called *adversarial attacks* can even undermine existing detection models because they are able to effectively conceal content on the basis of advance knowledge. Finally there are increasingly more platforms, like *Pastebin, Discord* or *Snapchat*, on which real-time information is being shared which is only temporarily available (volatile, or *ephemeral* content). The existence of such platforms motivates the development of new, *real time* detection techniques and methodologies of intervention.

Challenges in intervention on the basis of literature and practice

The public-private cooperation on online and offline interventions in the field of extremist content still finds itself in infancy phase. The focus for the government is mainly at the exploration of online interventions in collaboration structures, whereas the online platforms and service providers search for interventions aimed at specific target groups or problems. In addition the platforms want to do more on prevention of harmful content and its spreading than only on content moderation.
Public as well as private partners find it desirable that next to the more reactive approach of safety partners, a more integral approach will come in which social partners aiming at prevention are involved. Recently new organizations and cooperation partnerships have arisen which take on a more specific role in the public-private playing field on online extremism and radicalization. The ATKM in The Netherlands, GIFCT and Tech Against Terrorism international coordinate interventions, improve mutual processes, share knowledge, help (new) online platforms getting started and develop technology together. In addition new forms of cooperation from private and public organizations arise on their own. If these kind of cooperations are feasible and successful, is for a large part dependable on legislation, image, legitimacy and mutual trust. The lack of research on the effectiveness of online interventions on extremist and radical content necessitates to in any case more exchange of experiences and insights.

[2] Https://openai.com/blog/introducing-gpts

Inhoudsopgave

1 Inleiding

1.1 Behoeftestelling, onderzoeksvraag en afbakening

De minister van Justitie en Veiligheid heeft aan de Tweede Kamer toegezegd om medio 2023 te komen met een nadere strategie voor de aanpak van radicalisering en extremisme, met daarbij bijzondere aandacht voor preventie, mede als aanvulling op de implementatie van de EU Verordening 2021/784 inzake het tegengaan van terroristische online inhoud. Dit onderzoek is bedoeld om hieraan bij te dragen en beoogt met name het handelingsperspectief van de overheid te vergroten rond online radicalisering en extremisme. Daarbij biedt het onderzoek een overzicht van huidig gebruikte en state-of-the-art innovatieve technologische oplossingen t.b.v. detectie, monitoring en interventies om de verspreiding van online extremistische content tegen te gaan. Tevens biedt het technische en strategische beleidsaanbevelingen.

De onderzoeksvraag in dit project luidt:

Welke innovatieve technologische oplossingen zijn effectief voor a) de signalering/detectie en b) het tegengaan van de verspreiding van radicale en (rechts)extremistische content online en onder welke condities kunnen deze worden toegepast?

Bij aanvang van het onderzoek lag de focus op twee vormen van extremisme: rechts en anti-institutioneel online extremisme in de Nederlandse context. Tijdens ons onderzoek is anti-institutioneel extremisme als apart online fenomeen niet verder onderzocht, omdat het hier een vrij nieuw fenomeen betreft en nog weinig beschikbare literatuur kent.

1.2 Aanpak

Om de onderzoeksvraag te beantwoorden is een inventarisatie gemaakt van de state-of-the-art op het gebied van detectie en interventie. Hiervoor is gebruik gemaakt van interviews met publieke en private stakeholders gericht op *best practices* en behoeftestelling, heeft verdiepend literatuuronderzoek plaats gevonden en is een tweetal interactieve workshops met deze stakeholders georganiseerd waarin ook aanvullende behoeften vanuit de praktijk zijn geïnventariseerd.

1.2.1 Interviews

In afstemming met de NCTV zijn in totaal 17 interviews gehouden met stakeholders, waaronder overheidspartijen, uitvoeringsorganisaties en serviceproviders. In het kader van vertrouwelijkheid is besloten deze interviews anoniem te verwerken, op zowel persoonlijk als institutioneel vlak. In de interviews is getracht de beschikbare kennis bij stakeholders te inventariseren en zicht te krijgen op de huidige gehanteerde methodieken voor de detectie van online extremistische content, interventietechnieken en aanvullende behoeften. De interviews adresseerden het huidige overheidsbeleid omtrent online extremisme en radicalisering, de actuele praktijk van de betreffende stakeholder, de bekendheid met online radicalisering en extremisme, en implicaties van huidige trends in onderzoek en praktijk.

1.2.2 Literatuuronderzoek

Het literatuuronderzoek heeft de trends rond online extremisme en radicalisering op met name sociale mediaplatformen nader bekeken, met als aandachtspunten welke platformen voornamelijk gebruikt worden voor welke typen radicalisering en extremisme, op welke doelgroepen deze platformen zich richten, welke veranderingen ten aanzien van het (online) gedrag van gebruikers van deze platforms worden gezien, en wat bekend is over de effectiviteit van de gerapporteerde methodieken en technologieën voor de detectie van online extremistische content en radicalisering. Daarnaast heeft het literatuuronderzoek in kaart gebracht wat de huidige staat van detectietechnologie (gericht op tekst en beeld) binnen de gekozen scope is en wat de verwachte ontwikkelingen op dat vlak zijn.

1.2.3 Workshops

In totaal zijn er twee interactieve workshops georganiseerd met de eerder geïnterviewde stakeholders. Tijdens de eerste workshop is kort de stand van zaken van het onderzoek weergegeven en zijn de deelnemers in groepen uiteengegaan om aan de hand van prikkelende stellingen een blik te werpen op hun praktijk en behoeften. In deze *break-out sessies* werd gesproken over respectievelijk detectie- en interventiemethoden. Hierbij werd gevraagd naar de aanwezige kennis, het bestaande beleid, de effectiviteit en de behoeften van de organisaties.
Tijdens de tweede workshop richtte de sessie over detectie zich op technieken voor de detectie van extremisme en radicalisering in tekst en beeld. De sessie over interventie richtte zich op de verschillende interventies die vandaag en in de nabije toekomst mogelijk zijn, de huidige en verwachte effectiviteit daarvan, en prioriteiten op de korte en langere termijn.

Naast de inhoudelijke output betrof het doel van deze workshops ook het bij elkaar brengen van verschillende nationale en internationale stakeholders, zowel publiek als privaat. De groep was divers: Rijksoverheidspartijen, uitvoerende veiligheidsdiensten, grote sociale media platformen en een vertegenwoordiging vanuit serviceproviders.

1.2.4 Definities

Voor dit onderzoek zijn de volgende NCTV-definities uit het Dreigingsbeeld Terrorisme Nederland[3] aangehouden.

Extremisme
“Het uit ideologische motieven bereid zijn om niet-gewelddadige en/of gewelddadige activiteiten te verrichten die de democratische rechtsorde ondermijnen.”

Extreemrechts
“Tot extreemrechts gedachtegoed worden denkbeelden gerekend als vreemdelingenhaat (inclusief: antisemitisme en antimoslims), haat jegens vreemde (cultuur)elementen en ultranationalisme. Hierbij houden extreemrechtse activisten zich veelal aan de wet, overtreden rechts-extremisten de wet en extreemrechtse terroristen plegen op mensenlevens gericht geweld gebaseerd op extreemrechtse opvattingen.”

Anti-institutionalisme
"Anti-institutioneel extremisten blijven hun ‘kwaadaardig-elite-narratief’ uitdragen over een internationaal opererende elite die er op uit zou zijn om de bevolking te onderdrukken, tot

[3] Definities gebruikt in het Dreigingsbeeld Terrorisme Nederland | Dreigingsbeeld Terrorisme Nederland | Nationaal Coördinator Terrorismebestrijding en Veiligheid (nctv.nl)

slaaf te maken of zelfs te vermoorden. Daarnaast vormen zelfverklaarde soevereinen een prominente groep binnen de anti-institutionele beweging. Zij ontkennen de legitimiteit van de overheid en verklaren zichzelf onafhankelijk van de Nederlandse staat. In beide gevallen schuilt de dreiging met name in de ondermijning van de democratische rechtsorde."

Radicalisering
"Een proces van toenemende bereidheid om de uiterste consequentie uit een denkwijze te aanvaarden en die in daden om te zetten. Deze toenemende bereidheid kan leiden tot gedrag dat andere mensen diep kwetst of in hun vrijheid raakt, kan aanleiding zijn voor individuen of groepen om zich af te keren van de samenleving en kan leiden tot het gebruik van geweld."

Terrorisme
"Het uit ideologische motieven plegen van op mensenlevens gericht geweld, dan wel het aanrichten van maatschappij-ontwrichtende zaakschade, met als doel maatschappelijke ondermijning en destabilisatie te bewerkstelligen, de bevolking ernstige vrees aan te jagen of politieke besluitvorming te beïnvloeden."

Overige definities:

Serviceprovider
"Een serviceprovider is een bedrijf of organisatie die diensten levert aan bedrijven, individuen of organisaties. Deze diensten omvatten onder andere het aanbieden van internetdiensten, cloudopslag of chat services (Law Insider, 2024)."

Sociale media
Sociale media is een verzamelnaam voor alle online platformen en kanalen die interactie tussen de gebruikers mogelijk maken. Gebruikers op sociale media staan met elkaar in verbinding, bijvoorbeeld door online 'vrienden' met elkaar te worden en kunnen zo communiceren (Mediawijs, n.d.).

Overheidspartijen
Wanneer er in dit rapport gesproken wordt over overheidspartijen betreft het de NCTV en verschillende ketenpartners van de NCTV wat betreft het onderwerp online extremisme en radicalisering. Hieronder vallen ook de politie, gemeentes, en taakgroepen van verschillende ministeries. Hierbij moet worden opgemerkt dat NCTV geen rol heeft in zoekslagen met betrekking tot een casus-gerichte aanpak, en in die zin afwijkt van andere overheidspartijen.

Interventies
Acties die de verspreiding van online radicale en (rechts)extremistische content tegengaan.

Detectie
In de context van online radicalisering verwijst "detectie" naar het proces van het identificeren en opsporen van tekens, indicatoren of gedragingen die kunnen wijzen op het begin of de verspreiding van extremistische ideologieën, terroristische propaganda, of de radicalisering van individuen online.

1.2.5 Interviewprotocol

Voorafgaand aan de interviews hebben publieke en private partijen, waaronder beleidsorganisaties, handhavende en operationele diensten, sociale media en service providers, een uitnodiging toegezonden gekregen waarin de aanleiding en de opzet van het

onderzoek werd toegelicht. Indien gewenst is daar een nadere telefonische toelichting op gegeven. De genodigden hebben de afweging gemaakt om zelf aan het interview deel te nemen, dan wel een of meerdere collega's af te vaardigen. De interviewkandidaten zijn geselecteerd op basis van hun inhoudelijke kennis, in combinatie met hun ervaring op dit gebied binnen hun organisatie. De interviews zijn digitaal afgenomen aan de hand van een vragenlijst. Deze vragenlijst diende als leidraad voor de open gesprekken. Daardoor zijn niet in elk interview alle vragen even nadrukkelijk aan bod gekomen. In totaal zijn 16 interviews afgenomen. De vragen hadden betrekking op de grondslagen om te detecteren en te interveniëren om verspreiding te voorkomen, welke doelstellingen de organisatie kent op dit domein en welke taken zij heeft, maar ook de verandering die zij daarin heeft waargenomen, welke trends noemenswaardig zijn, of het onderwerp prioriteit voor de organisatie heeft en tot slot waar de behoeften liggen.

1.2.6 Wegwijzer en voorbehoud

Deze studie benadert de fenomenen online extremisme en radicalisering vanuit een zowel theoretisch als praktisch, toegepast perspectief (Hoofdstuk 4), en richt zich op daarbij op detectie (Hoofdstuk 2) en interventie (Hoofdstuk 3). De diverse kruisverbanden en relaties tussen deze onderwerpen worden benoemd, maar in deze studie binnen de gekozen scope niet verder uitgediept. De hoofdstukken 2 en 3 kunnen dan ook los van elkaar gelezen worden. In Hoofdstuk 5 (Conclusies) benoemen we interessante onderwerpen voor vervolgonderzoek gericht op het verder onderzoeken van de connectie tussen de detectie van en interventie bij online extremisme en radicalisering.

2 Factoren in radicalisering en trends in online extremisme

Dit hoofdstuk beschrijft een beknopt literatuuroverzicht naar online extremisme en radicalisering alsmede de resultaten van gesprekken met stakeholders. Er wordt ingegaan op factoren die een rol spelen in het radicaliseringsproces, en het verschil tussen offline en online radicalisering wordt geduid. Vervolgens wordt een aantal trends beschreven in online extremisme en radicalisering.

2.1.1 Factoren in radicalisering

De afgelopen decennia is veel onderzoek gedaan naar individuele radicaliseringsprocessen en factoren die daarin een rol spelen. Radicalisering blijkt te worden beïnvloed door een veelvoud aan factoren, waarbij elk individueel radicaliseringsproces er anders uitziet. Desalniettemin is een aantal categorieën van relevante factoren te onderscheiden.

(Nickolson, Bergen, Feddes, Mann, & Doosje, 2021) zeggen daarover:

"Onze analyse van deze empirische onderzoeken laat zien dat radicalisering tot extremisme een complex proces is. Er is niet één factor die ontvankelijkheid voor extremisme of extremistisch handelen veroorzaakt. De empirische onderzoeken vinden doorgaans namelijk niet dergelijke causale verbanden, maar kunnen hoogstens aantonen welke factoren de kans op extremisme vergroten. Deze kans neemt toe wanneer er sprake is van een combinatie of accumulatie van factoren."

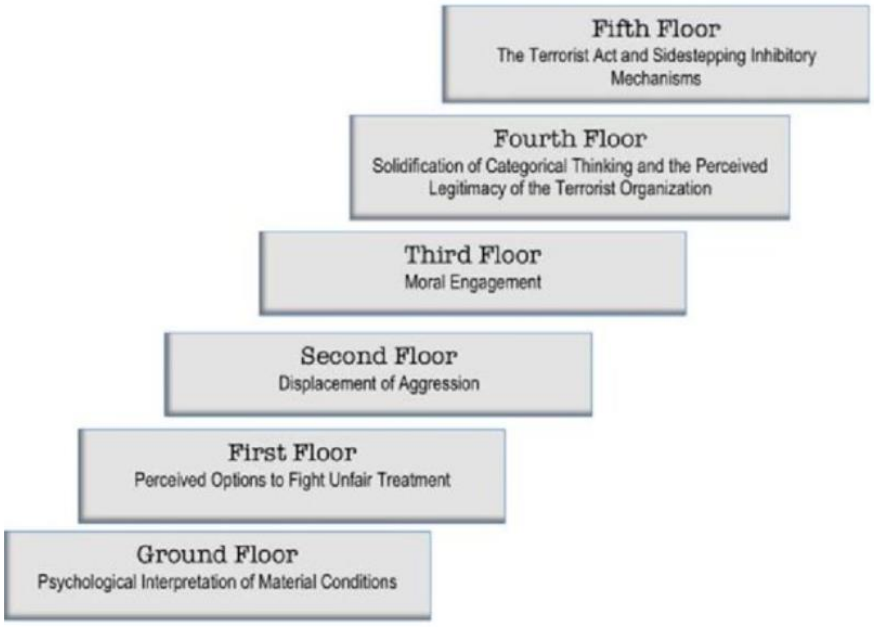
Onder veel genoemde demografische variabelen die kunnen bijdragen aan het proces van radicalisering vallen leeftijd, sociaaleconomische status, eerdere arrestaties, opleiding, werk, relatiestatus en de woonwijk (Desmarais, Simons-Rudolph, Brugh, Schilling, & Hoggan, 2017). Zo laten studies zien dat hoewel het percentage vrouwen die radicaliseren toeneemt, mannen meer deelnemen aan gewelddadige extremistische groeperingen (Cunningham, 2007). Vrouwen zijn aanwezig in extremistische bewegingen maar nemen daarin vaak een achtergrondrol in, zoals in het werven van nieuwe leden. Binnen extreemrechtse bewegingen bijvoorbeeld worden vrouwen ingezet om mannen te werven door het gevoel van mannelijke superioriteit te bevestigen (Leidig E. , 2021). Wat leeftijd betreft wordt aangenomen dat mensen in hun adolescentie het meest kwetsbaar zijn voor radicalisering en extremisme (Bhui, Dinos, & Jones, 2012). De relatie met radicalisering en extremisme is soms ambigu; zo houdt zowel een hoog als een laag opleidingsniveau verband met een verhoogde kwetsbaarheid voor radicalisering (Schmid A. P., 2013). Er kan ook een interactie zijn tussen demografische factoren; mannen nemen bijvoorbeeld vaker dan vrouwen hun toevlucht tot delinquentie wanneer zij worden geconfronteerd met stress in hun persoonlijke omgeving (Silke, 2008).

Daarnaast wordt een aantal psychologische variabelen geassocieerd met extremisme. Gevoelens van relatieve deprivatie van een sociale groep, een gevoel van onrechtvaardigheid, discriminatie, ongelijkheid, marginalisering, wrok, sociale uitsluiting, frustratie, slachtofferschap, stigmatisering, banden met extremistische groeperingen en een zoektocht naar sociale identiteit en betekenis worden in dit verband genoemd (Bond, 2014) (Kruglanski, et al., 2014) (Piazza, 2006) (Sageman, 2004) (Vergani, Iqbal, Ilbahar, & Barton, 2018). In het bijzonder behoeften (*needs*), en de mate waarin die al dan niet vervuld worden, kunnen mensen motiveren zich te verbinden aan extremistische denkbeelden en groepen. Voorbeelden van behoeften zijn de behoefte aan rechtvaardigheid, de behoefte aan identiteit, de behoefte aan verbondenheid, de behoefte aan betekenis en de behoefte aan sensatie (Borum, 2012) (Feddes, Nickolson, & Doosje, 2015) (Kruglanski, Violent radicalism and the psychology of prepossession, 2018).

(Sterkenburg, 2021) onderscheidt vijf typen extremisten binnen de extreemrechtse beweging in Nederland: rechtvaardigheidszoekers zijn boos op de overheid, politieke zoekers zijn op zoek naar bredere steun, spanningzoekers willen aanvankelijk vooral provoceren, sociale zoekers zijn op zoek naar vriendschap of willen bestaande relaties bestendigen, en ideologische zoekers zien hun deelname vooral als een ideologische queeste en ultieme zelfverwezenlijking. Hoewel deze activisten een andere reden hebben voor hun extremistische opvattingen komen zij elkaar wel vaak tegen binnen de bewegingen.

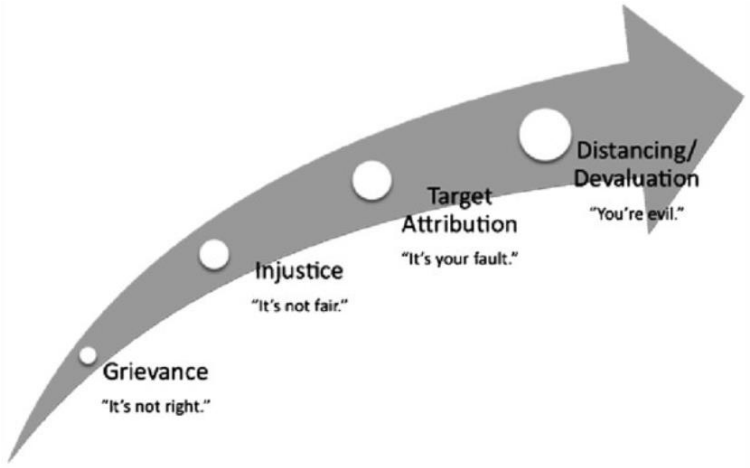
2.1.2 Radicaliseringsmodellen

Verschillende (psychologische) modellen beschrijven het proces richting extremisme. Een bekend voorbeeld is de trap naar terrorisme (*staircase toward terrorism*) van (Moghaddam, 2005), waarin individuele radicalisering zich ontwikkelt vanuit ervaren onrechtvaardigheid in iemands omstandigheden, via het opbouwen van woede en frustratie, tot de bereidheid om geweld te gebruiken (zie Figuur 2.1).



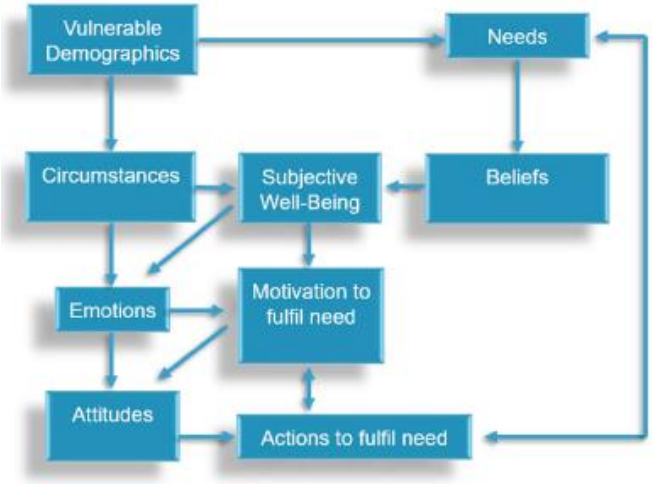
Figuur 2.1: Het Staircase to Terrorism model (Moghaddam, 2005).

Ook Borum's model (2012; zie Figuur 2.2) gaat uit van een model met stadia van toenemende ernst, met grieven (*het is niet juist*), gevoelens van onrechtvaardigheid (*het is niet eerlijk*), anderen de schuld geven (*het is jouw schuld*), tot afstand nemen en devaluatie (*jij bent slecht)*.



Figuur 2.2: Het Four-Stage Model of the Terrorist Mindset (Borum, 2012).

Een meer procesmatig model wordt voorgesteld door (Van den Berg & Van Hemert, 2021). Hun General Needs and Affect (GNA) process model voor gewelddadig extremisme (zie Figuur 2.3) beschrijft radicalisering als een proces waarbij individuen op basis van demografische kenmerken en omstandigheden, de daaraan gekoppelde behoeften en de (ervaren) bevrediging daarvan kunnen worden aangezet tot acties.



Figuur 2.3: Het General Needs and Affect (GNA) process model voor gewelddadig extremisme (Van den Berg & Van Hemert, 2021).

2.1.3 Offline versus online radicalisering

Er bestaan binnen wetenschappelijke literatuur verschillende perspectieven op de rol van het internet bij radicalisering en extremisme. Eén van die perspectieven is dat online radicalisering een alternatief is voor offline radicalisering, waarbij het proces voor online radicalisering hetzelfde verloopt als voor offline radicalisering en extremistisch gedrag. Een

contrasterend perspectief ziet online radicalisering als een volledig nieuw proces dat zich alleen online manifesteert.

Het scheiden van offline en online werelden heeft echter weinig zin; technologische innovatie heeft de twee onlosmakelijk met elkaar verbonden (Whittaker J. , 2023). Er zijn immers maar weinig mensen die uitsluitend internet gebruiken of helemaal geen gebruik maken van internet (Herath & Whittaker, 2023). Bovendien zijn verschillende rollen (bijvoorbeeld financiën, planning, bommen maken, reizen, netwerken) op verschillende manieren afhankelijk van internet.

Over het algemeen wordt er geen causale relatie gevonden tussen internetgebruik en radicalisering (Whittaker J. , 2022). Online activiteiten kunnen echter wel een risicofactor zijn voor het radicaliseren. (Van Wonderen, et al., 2023) beargumenteren dat social media weliswaar een beperkte rol hebben in radicalisering, maar een grote rol in de normalisering van extremistisch gedachtegoed.
In een poging om de dichotomie tussen online en offline radicalisering te overstijgen, identificeren (Herath & Whittaker, 2023) vier paden naar radicalisering op basis van enkele honderden casussen van Islamic State (IS) terroristen:

1. Het 'geïntegreerde' traject met een hoge netwerkbetrokkenheid, zowel online als offline, meestal bestaande uit individuen die als onderdeel van een groep acties plannen en voorbereiden
2. Het 'aangemoedigde' traject omvat individuen die meer in het online domein handelen ten koste van het offline domein.
3. Terroristen in het 'geïsoleerde' traject worden gekenmerkt door een gebrek aan interactie tussen de offline en online domeinen.
4. Het 'ingesloten' traject omvat actoren die meer offline netwerkactiviteit vertonen, maar nog steeds het internet gebruikt voor het plannen van hun activiteiten.

2.1.4 Trends in online radicalisering en extremisme

Op basis van zowel de literatuur als de interviews met stakeholders onderscheiden we een aantal trends in manieren van opereren op het gebied van online radicalisering en extremisme.

In het algemeen is online radicalisering aan verandering onderhevig (NCTV, Dreigingsbeeld Terrorisme Nederland, 2023). Zo wordt communicatie onder gelijkgestemden steeds internationaler en steeds heimelijker. Daarnaast gaat het *ontvangen* van propaganda steeds meer gepaard met het *produceren en verspreiden* van propaganda. De aard van de online radicalisering wordt diverser; extremistische denkbeelden uit verschillende hoeken worden gekoppeld aan geweldsfantasieën complottheorieën. Tenslotte wordt online radicalisering vaker gezien bij jonge mensen en zelfs minderjarigen, omdat leeftijdsverschillen, die in de fysieke wereld een obstakel kunnen zijn voor contact, in de online wereld geen belemmering vormen. Het generieke beeld is dus dat online radicalisering sneller, internationaler, diverser en jonger wordt.

Een belangrijke trend betreft het gebruiken van populaire en actuele onderwerpen om extremistisch gedachtengoed te introduceren. Extremistische groepen doen dit bijvoorbeeld door middel van *red pilling:* het creëren van online content rondom actuele en populaire onderwerpen en die onderwerpen vervolgens inzetten om extremistisch gedachtegoed te verspreiden.

Een tweede gerelateerde trend is het normaliseren van extremistisch gedachtengoed. Door het injecteren van extremistische denkbeelden op reguliere platformen in relatie tot mainstream onderwerpen groeit de acceptatie van deze denkbeelden in de samenleving. Voorbeelden van onderwerpen die actief en bewust op deze manier worden gebruikt zijn fitness en *cottage core*[4], het benadrukken van respectievelijk lichamelijke fitheid en het geromantiseerde buitenleven. Hierdoor zijn groepen als de *incel* of *tradwife* communities[5][6] de laatste jaren sterk gegroeid (Dodd, 2023) (Leidig E. , 2021). Het gebruik van influencers, bijvoorbeeld op het gebied van lifestyle en fitness, helpt bij deze normalisering (Van Wonderen, et al., 2023). Een voorbeeld is de rol van vrouwen in zogenaamde Active Clubs. Active Clubs zijn extreemrechtse groeperingen die aan mentale en fysieke training doen om zich voor te bereiden op een eventuele rassenoorlog. Ze zijn tot nu toe voornamelijk aanwezig in de Verenigde Staten en Canada. Active Clubs gebruiken accounts en afbeeldingen van (witte) vrouwen die in eerste instantie de nadruk leggen op fitness en zelfstandigheid van vrouwen, een beeld dat afwijkt van de meer traditionele rol van vrouwen in de extreemrechtse beweging. Dit lijkt een strategie om meer vrouwen aan te trekken zonder hen te beperken tot huwelijk en moederschap. Hoewel de beelden mogelijk aantrekkelijk zijn voor geïsoleerde vrouwen blijven traditioneel gendergedrag en vrouwenhaat aanwezig. De normalisering van extremistisch gedachtegoed wordt niet alleen gezien in het adopteren van 'mainstream trends' zoals de genoemde *cottage core* of *tradwives*, maar ook door het infiltreren van bestaande communities zoals in gaming (Wells, et al., 2023). Doordat er een platform wordt gegeven aan extremistische uitingen wordt het idee gecreëerd dat deze uitingen acceptabel zijn en zullen meer mensen zich aangesproken voelen (Bolet & Foos, 2023).

Een derde trend is het groeiende besef dat zogenaamde aanbevelingsalgoritmes op social mediaplatforms niet *alleen* verantwoordelijk zijn voor online radicalisering en extremisme. Het beeld van argeloze internetgebruikers die in een extremistische fuik worden getrokken en daardoor min of meer automatisch radicaliseren wordt in een rapport van (Van Wonderen, et al., 2023) genuanceerd. Minstens zo belangrijk als aanbevelingsalgoritmes op social mediaplatforms zijn de persoonlijke keuzes die gebruikers maken, zoals het delen van links naar extremistische content, het volgen van bepaalde kanalen op sociale mediaplatforms en persoonlijk zoekgedrag. Deze bevinding bevestigt het belang van onderzoek naar menselijke biases, zoals de neiging om steeds extremere content te willen zien en de neiging om bevestiging voor bestaande overtuigingen te zoeken (*confirmation bias*). De platforms investeren bovendien meer capaciteit in moderatie van en onderzoek naar extremistische content, waarbij ook de aanpassing van algoritmen in acht wordt genomen.

Een vierde trend wordt ook genoemd door (Van Wonderen, et al., 2023), en betreft polariserende borderline content, die weliswaar legaal is maar wel schadelijk kan zijn. Dit gaat onder andere om *dog whistles* (hondenfluitjes) en het gebruik van woorden die net anders zijn dan de woorden waar de moderatiealgoritmes van social mediaplatforms op zijn ingericht, maar wel een soortgelijke interpretatie of associatie oproepen. Dit soort *dog whistles*[7] maakt automatisering van detectie en moderatie lastig. Een voorbeeld hiervan is de "Boogaloo"-beweging, die door Meta in 2020 werd verwijderd van Facebook en

[4] Cottagecore - Wikipedia
[5] 'Tradwives': the new trend for submissive women has a dark heart and history | Fashion | The Guardian
[6] De incel-community bestaat uit individuen die beweren geen romantische of seksuele partner te kunnen vinden, hiervan de schuld bij de maatschappij en/of het feminisme leggen, en waarvan sommigen extreme en vaak misogyne opvattingen uiten. Tradwives zijn vrouwen die zich identificeren met een traditionele, huiselijke rol en streven naar een levensstijl waarin zij zich voornamelijk richten op huishoudelijke taken en het ondersteunen van hun partner.
[7] Dog whistle (politics) - Wikipedia

Instagram.[8] Deze beweging, oorspronkelijk uit de Verenigde Staten, is een los georganiseerde, rechtse, anti-overheidsbeweging die bekendstaat om haar verzet tegen vermeende overheidsbemoeienis en haar focus op het recht op wapenbezit. Hoewel de initiële verbanning succesvol was, werd duidelijk dat de hoeveelheid Boogaloo-content op het web na verbanning hetzelfde was gebleven. De online berichten waren niet meer dezelfde, maar de onderliggende boodschappen wel. Zo werd er niet meer gesproken over Boogaloo maar over 'big igloo' of 'the luau', of de naam werd vervangen door het iglo icoon. Daarnaast werden afbeeldingen gebruikt die geassocieerd worden met de beweging, zoals Hawaiiaanse shirts en de *Shiba Inu Doge* meme (een afbeelding van een hond met teksten in gebroken Engels). Deze bleven terug komen op verschillende platformen en op Facebook zelf. Meta's model was destijds getraind op taal, niet op afbeeldingen, waardoor het moeilijker werd om dergelijke versluierde berichten te detecteren. De tekst rondom de gebruikte afbeeldingen bleef echter hetzelfde, waardoor het nog steeds mogelijk was om bepaalde memes te vinden. Deze casus van Boogaloo benadrukt het belang van genuanceerde modellen om de strategische omzeiling van platformbeleid te voorkomen.

Een laatste trend is het migreren tussen platforms en service providers, zowel door extremistische individuen als van content. Gebruikmakend van de boven beschreven tactieken blijft extremistische content geplaatst worden op grotere platformen waar veel verkeer is. Wanneer een gebruiker interesse toont in het onderwerp wordt de discussie verplaatst naar een kleiner platform, waar de middelen voor moderatie beperkter zijn, of de welwillendheid tot moderatie vanuit het platform beperkt of zelfs afwezig is. Zo wordt er geworven op grote platforms, maar gebeurt de daadwerkelijke radicalisering op de kleinere platforms waar minder toezicht op is.

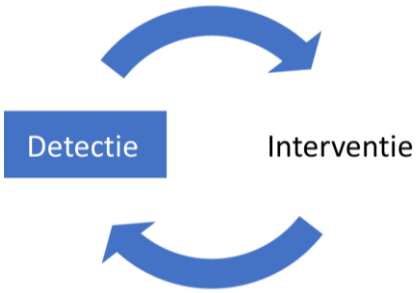
2.2 Conclusie

In dit hoofdstuk zijn we op beknopte wijze ingegaan op trends in online extremisme en radicalisering, zoals gerapporteerd in recente literatuur. We zijn ingegaan op factoren die een rol spelen bij deze fenomenen, hebben radicaliseringsmodellen besproken, en hebben een kort overzicht gepresenteerd van recente trends rond online extremisme en radicalisering. De huidige trends op het gebied van online radicalisering en extremisme zijn in lijnen met reguliere modellen uit de radicaliseringsliteratuur, in de zin dat stadia en fasen in radicalisering nog steeds op dezelfde manier doorlopen kunnen worden. Het verschil tussen online en offline radicalisering zit eerder in de omvang; gebruik van bijvoorbeeld mainstream onderwerpen (door *red pilling* en influencers) om mensen bij een groep te betrekken is niet in tegenspraak met modellen voor offline radicalisering, maar vindt wel plaats op grotere schaal dan bijvoorbeeld het organiseren van fysieke bijeenkomsten en netwerken waar zoekenden zich gezien voelen en langzaam in extremistische denkbeelden worden meegenomen.

[8] Banning a Violent Network in the US | Meta (fb.com)

3 Detectie van online radicalisering en extremisme

In dit hoofdstuk wordt ingegaan op een selectie van interessante bevindingen met betrekking tot recente (technische) mogelijkheden en hun beperkingen op het gebied van detectie van radicale en extremistische content, zoals die naar voren zijn gekomen uit enerzijds (wetenschappelijke) literatuur en anderzijds uit de praktijk. De status uit de praktijk van zowel overheidspartijen als bedrijven is voortgekomen uit interviews.



We beschrijven hier een aantal state-of-the-art methoden voor het detecteren van online radicalisering en extremisme in tekst en op sociale media. Deze methoden variëren van het beoordelen van extremistische en radicale inhoud in een tekst tot het analyseren van trends in radicalisering en extremisme of het opsporen van geradicaliseerde personen. De meeste state-of-the-art methoden vallen onder de noemer *Natural Language Processing (NLP)*: een verzamelnaam voor allerlei technieken om tekst te laten verwerken door computers. In dit geval worden NLP-technieken ingezet om teksten van sociale media te classificeren met als doel om extremisme op te sporen. Dit wordt gedaan met Machine Learning-technieken en Deep Learning-technieken. Deze methoden worden toegelicht in Bijlage A. Tenslotte wordt een techniek beschreven die niet binnen NLP valt, namelijk Social Network Analysis: de analyse van netwerken van, in dit geval, extremistische en geradicaliseerde gebruikers op sociale media.

In elke subsectie van dit hoofdstuk wordt de desbetreffende methode kort beschreven en worden voorbeelden gegeven van technieken die op een succesvolle manier radicalisering en extremisme opsporen op sociale media. Bij de selectie van deze technieken is gelet op de door de onderzoekers gerapporteerde performance scores – hoe succesvol het betreffende model is op onderzoeksdata – en de eventuele toepasbaarheid: de methode moet niet alleen theoretisch interessant zijn maar ook praktisch nut hebben voor stakeholders.

3.1 Detectiemethodieken uit de literatuur

Door te kijken naar teksten die geschreven zijn door geradicaliseerde en extremistische personen kunnen bepaalde tekstuele eigenschappen ("features", in machine learning-jargon) worden onderscheiden die in deze teksten bovengemiddeld vaak voorkomen. Als die eigenschappen zich laten uitdrukken in tekstuele features en voldoende leerbaar zijn, dan kan een machine learning-model voorspellen welke teksten door geradicaliseerde personen zijn geschreven en welke niet. Features die in deze context vaak gebruikt worden zijn *n*-grams (Jaki & De Smedt, 2018; Oussalah et al., 2018, De Smedt et al. 2018, Mashechkin et al., 2019), het gebruik van bepaalde woorden (Rowe & Saif, 2016), de lengte van de tekst, emoji en leestekens (Oussalah et al., 2018), hashtags (Ashcroft et al., 2015) en de uitgedrukte emoties in de tekst (Araque & Iglesias, 2019).

We lichten twee voorbeelden uit van papers die gebruik maken van machine learning-technieken om te bepalen of een tekst van een geradicaliseerd/extremistisch persoon komt. Deze artikelen worden gekenmerkt door hun vernieuwende technieken en hun gerapporteerde hoge nauwkeurigheid.

3.1.1 Radicalisering opsporen door middel van woorden, psychologie en gedrag

Een voorbeeld van het gebruik van lexicale (woordgebaseerde) detectiemethoden voor online extremisme is de ontwikkeling van de Violence Risk Index door Ebner, Kavanagh en Whitehouse (Julia Ebner, 2023). Deze studie toont aan dat bepaalde taalkundige indicatoren gelinkt kunnen worden aan de bereidheid om gebruik te maken van geweld, en dat die indicatoren aangetoond kunnen worden in online groepen. Een dergelijke risicoanalyse kan gebruikt worden door bijvoorbeeld veiligheidsdiensten maar ook door *serviceproviders*.

Nouh, Nurse en Goldsmith (2019) hebben als doel om radicalisering in tweets over IS op te sporen en gebruiken hiervoor drie soorten features:

1. Radicaal taalgebruik,
2. Psychologische signalen,
3. Gedragssignalen.

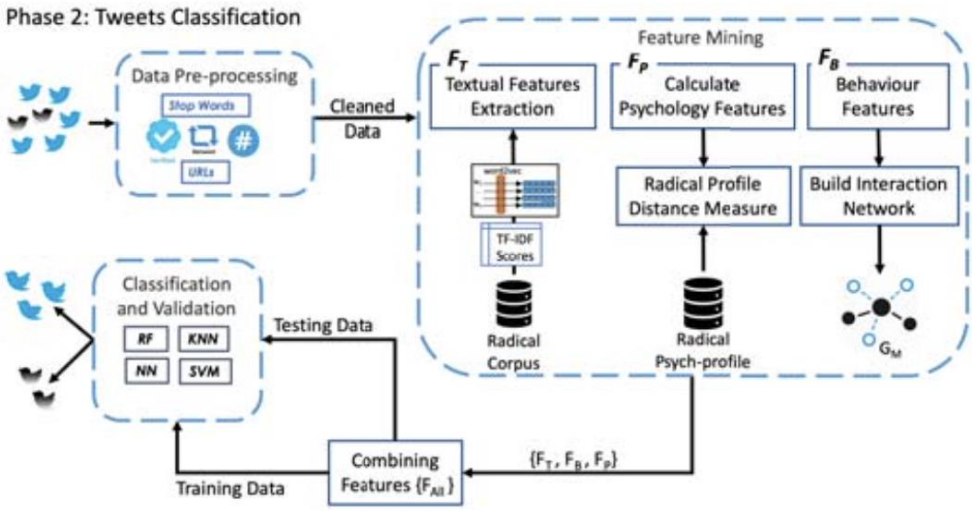
Om te bepalen welke tweets radicaal taalgebruik bevatten, vergelijken de auteurs het taalgebruik in de tweets met het taalgebruik van IS zelf in hun online tijdschrift *Dabiq*. Ze kijken hierbij naar hoe vaak bepaalde woorden voorkomen in *Dabiq* en in de tweets met de hypothese dat hoe vaker veelvoorkomende woorden uit *Dabiq* in een tweet voorkomen, hoe groter de kans is dat de tweet komt van een geradicaliseerd persoon.[9]

Voor het gebruiken van psychologische signalen als features gebruiken Nouh, Nurse en Goldsmith de Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Boyd, Jordan, & Blackburn, 2015). Op basis van het taalgebruik van de twittergebruiker wordt een psychologisch profiel gebouwd met factoren als positieve en negatieve emotie, karaktereigenschappen zoals extraversie en openheid en persoonlijke drijfveren zoals macht vergaren en horen bij een groep. Al deze factoren worden als features toegevoegd aan het model.

[9] De methode is in feite wat complexer en maakt ook gebruik van zogenaamde *embeddings*: numerieke representaties van woorden die worden gebruikt om betekenis te weergeven.

De gedragssignalen die worden gebruikt hebben te maken met hoe de gebruiker zich op Twitter gedraagt. Hierbij wordt gekeken naar onder andere hoe actief de gebruiker is, hoeveel volgers zij heeft en hoeveel mensen ze zelf volgt en de centraliteit (zie de paragraaf over Social Network Analysis in dit rapport) van de gebruiker in haar netwerk.

In Figuur 3.1 is te zien hoe dit in zijn werk gaat. De data krijgen, na te zijn opgeschoond, drie soorteneigenschappen ("features") toegekend: tekst, psychologie en gedrag. Hierna worden de features gecombineerd en wordt een model getraind en getest met deze features.[10]



Figuur 3.1: Methodiek van Nouh et al. (2019).

Na het invoeren van deze features trainen en testen de auteurs een model op drie datasets:
1. een dataset met tweets van geradicaliseerde IS-volgers (met als verificatiemethode een check of Twitter deze accounts verwijderd heeft wegens het breken van de regels);
2. een dataset met tweets over willekeurige onderwerpen die niets met IS te maken hebben;
3. een dataset met tweets over IS door niet-geradicaliseerde gebruikers.

Het model slaagt er niet alleen in om tweets door geradicaliseerde IS-volgers te onderscheiden van tweets over willekeurige onderwerpen maar ook om deze te onderscheiden van tweets over IS door niet-geradicaliseerde Twittergebruikers. Dit doet het model met een F-score van 1, de hoogst mogelijke score. Hierbij zijn de psychologische features het meest doorslaggevend: ook zonder de gedragsfeatures en taalfeatures, met alleen psychologische features, haalt het model een perfecte score.

Deze methode kan dus gebruikt worden voor het opsporen van geradicaliseerde personen op sociale media en kan daardoor nuttig zijn voor overheidsinstanties die radicalisering op Twitter willen monitoren.

[10] RF, KNN, NN en SVM zijn afkortingen van verschillende modellen die de auteurs hebben gebruikt.

3.1.2 Geradicaliseerde personen volgen in de tijd

Klausen, Marks, & Zaman (2018) hebben een zeer uitgebreide methode ontwikkeld om geradicaliseerde personen op te sporen en te volgen in de tijd. Hun onderzoek is gebaseerd op de observatie dat geradicaliseerde personen vaak geschorst worden van Twitter [11] en vervolgens opnieuw opduiken met een andere account. Het doel van dit werk is om deze personen op te kunnen sporen ook als ze van account wisselen. De auteurs gebruiken voor hun onderzoek een grote Twitter-dataset met accounts van IS-volgers.

De eerste stap in hun werk is het opsporen van geradicaliseerde personen. Belangrijke indicatoren zijn hiervoor gedragsfeatures, zoals de vraag of een bepaald account een bekende IS'er volgt, wanneer het account gecreëerd is, de hoeveelheid connecties en de hoeveelheid tweets van het account, of de geolocatie van het account aanstaat, of de tweets van het account publiek of privé zijn en of het account geverifieerd is door Twitter. De auteurs definiëren een geradicaliseerde gebruiker als een gebruiker die wordt geschorst door Twitter (dit omdat schorsing in de gevallen van accounts gelieerd aan IS vrijwel altijd wordt veroorzaakt door het posten van gewelddadig materiaal). Het doel is dus om door middel van de hierboven beschreven features te voorspellen of een account na verloop van tijd geschorst wordt. Uit de experimenten van de onderzoekers blijkt dat dit lukt met een nauwkeurigheidscore van 0,83.[12]

Klausen *et al.* (2018) constateren dat wanneer geradicaliseerde Twittergebruikers geschorst worden, ze vaak opnieuw verschijnen met een ander account. Daarnaast observeren ze dat het nieuwe account meestal grote gelijkenissen vertoont met het vorige account.[13] Deze informatie gebruiken deze auteurs om, door middel van oude, geschorste accounts, deze nieuwe accounts op te sporen. De features die ze hiervoor gebruiken zijn de gebruikersnaam en naam, de profielfoto en de banner van de account. Met deze methode kunnen ze zeer effectief, met een nauwkeurigheid van 0,91, voorspellen of twee accounts bij dezelfde persoon horen.

In Tabel 3.1 zijn twee accounts te zien die vrijwel gelijk zijn. De gelijkenis tussen de naam, foto en banner, gemeten in de laatste kolom φ, is volledig, en de gebruikersnamen komen overeen met een score van 0,88. Het feit dat dit model zo goed werkt komt grotendeels doordat geschorste Twittergebruikers vaak een nieuwe account openen met kenmerken die vrijwel identiek zijn aan die van hun oude account.

Tabel 3.1: Gelijkenissen tussen twee Twittergebruikers (uit Klausen et al., 2018).

Feature (k)	User i	User j	$\phi_k^{(i,j)}$
User ID	3307258107	3297609231	[NA]
Screen Name	Ahmes_Zirve__	Ahmes__Zirve	0.88
Name	Ahmes Zirve	Ahmes Zirve	1.00
Profile Picture	ff...ff	ff...ff	1.00
Profile Banner	[None]	[None]	1.00

[11] The former name of the messaging platform currently known as X.
[12] De auteurs van dit paper hebben geen F-scores gebruikt maar zogenaamde AUC-scores. Een AUC-score geeft weer hoe goed een model onderscheid kan maken tussen positieve en negatieve voorbeelden en de schaal loopt, net als F-scores, van 0 tot 1.
[13] Gelijkenissen tussen twee accounts worden berekend met de zogenaamde *Levenshtein distance*: een methode om de gelijkenis tussen twee teksten te berekenen.

Daarnaast kijken de auteurs naar de accounts die een bepaalde gebruiker volgt. De kans is groot dat als een gebruiker geschorst wordt en een nieuw account aanmaakt, deze gebruiker een aantal accounts zal gaan volgen die hij al volgde. Op basis van deze informatie kunnen de auteurs met een nauwkeurigheid van 0,66 voorspellen of twee accounts bij dezelfde persoon horen.

Een alternatieve, geavanceerdere manier om accounts te vergelijken richt zich op het analyseren van de tekstuele boodschappen die een account verspreidt. Stylometrische analyse kan worden toegepast om de gelijkenis tussen verschillende teksten te berekenen, waarmee een vorm van auteurschapstoekenning wordt geïmplementeerd. Het werk van Ranaldi *et al.* (2022) past deep learning-gebaseerde NLP-technieken toe om dark web accounts toe te kennen aan auteurs.

Een laatste bijdrage van dit paper is dat de auteurs een manier hebben ontwikkeld om efficiënt te zoeken naar nieuwe twittergebruikers. Hierdoor is het voor overheidsinstanties mogelijk om bij te houden welke nieuwe twittergebruikers erbij komen en of die nieuwe accounts bij dezelfde geradicaliseerde personen horen die eerder geschorst zijn. Dit maakt het vinden en vooral ook het volgen van geradicaliseerde personen makkelijker en efficiënter.

Naast de wat traditionelere Machine Learning-technieken worden de nieuwere *deep learning*-technieken de afgelopen jaren ook veelvuldig gebruikt in het domein van Natural Language Processing.[14] De afgelopen jaren zijn op taal gerichte deep learning-modellen ook ingezet voor het opsporen van radicalisering op sociale media. Veel gebruikte modellen zijn zogenaamde convolutional neural networks (*CNNs)* (Ahmad, Zubair Asghar, Alotaibe, & Awan (2019), Theodosiadou et al. (2021)), *LSTMs* (Ahmad et al. (2019), Kaur, Kaur Saini, & Bansal (2019), Rajendran et al. (2022)) en *BERT*-modellen (Jamil, Luqman, Pais, Cordeiro, & Dias (2022), Rajendran et al. (2022)).

Hieronder bespreken we twee andere methoden die interessante resultaten opleveren en goed scoren.

3.1.3 Terrorisme-gerelateerde content opsporen en volgen in de tijd

Het werk van Theodosiadou et al. (2021) combineert twee verschillende technieken: CNNs (Convolutional Neural Networks) en CPD (Change Point Detection). [15] Met deze technieken bepalen de auteurs met een hoge betrouwbaarheid wanneer een Twitter-post (tweet) terrorisme gerelateerd of haat zaaiend is of niet en detecteren zij wanneer er een significante verandering in de hoeveelheid van dit soort teksten plaatsvindt in de tijd. Alhoewel dit werk zich richt op tweets die terrorisme gerelateerde en haat zaaiende teksten bevatten en dus minder gericht is op radicalisering, kan de techniek die zij hebben ontwikkeld ook in dat kader nuttig zijn.

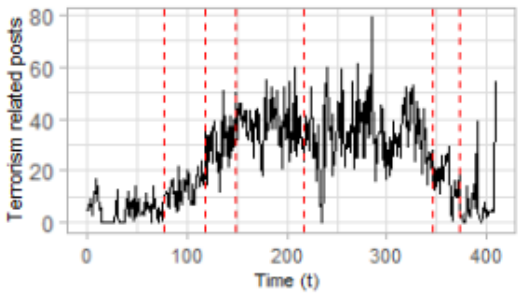
Ten eerste kunnen historische Twitterdata gebruikt worden om te bepalen wanneer er significante veranderingen plaats hebben gevonden in de mate van radicalisering onder

[14] Zie Bijlage A voor nadere informatie over Deep Learning.
[15] Een CNN is een soort neuraal netwerk dat met name gebruikt wordt in het domein van beeldherkenning. Dit komt doordat het model een laag bevat dat steeds kleine stukjes van het beeld interpreteert, waardoor het steeds beter wordt in het herkennen van structuren. Hoewel deze intuïtie niet direct te vertalen is naar het NLP-domein, blijken CNNs ook goed te werken als taalmodellen. Met Change Point Detection kunnen radicale veranderingen in tijdsreeksen gedetecteerd worden.

internetgebruikers. Deze veranderingen kunnen gekoppeld worden aan andere gebeurtenissen rond deze tijd (denk hierbij bijvoorbeeld aan bepaalde maatregelen in de Coronatijd) en inzicht geven in hoe bepaalde radicaliserende burgers reageren op wat er in de maatschappij gebeurt. Ten tweede kan in real-time worden gekeken of de radicalisering van personen op sociale media significant toe- of afneemt. Op basis van deze informatie kan beleid worden aangepast en kunnen partijen binnen het veiligheids- en justitiedomein extra alert zijn op gevaarlijke ontwikkelingen.

De auteurs van dit paper maken hier gebruik van door een CNN te trainen op twee taken: het herkennen van haat zaaiende teksten en het herkennen van terrorisme gerelateerd materiaal in Twitterdatasets. De CNN kan zowel de haat zaaiende als de terrorisme gerelateerde tweets correct classificeren met een F-score van 0,93. Daarnaast maken de onderzoekers gebruik van het feit dat elke tweet in de dataset een tijdsindicatie heeft. Je kunt dus zien wanneer er meer of juist minder terrorisme gerelateerde tweets worden geplaatst. In dit geval wordt er berekend wanneer er een patroondoorbreking is in de hoeveelheid tweets en naar aanleiding van dit soort patroondoorbrekingen worden de tweets ingedeeld in tijdsvakken. In Figuur 3.2 is de hoeveelheid terrorisme gerelateerde tweets uit de dataset van Theodosiadou et al. weergegeven. We zien dat de hoeveelheid sterk verandert in de tijd en dat die veranderingen te kenmerken zijn door zes patroononderbrekingen die de tweets in zeven tijdsvakken verdelen.

De auteurs merken op dat de patroononderbrekingen in hun data vaak samengaan met terroristische daden. Zij zien met name dat er na terroristische daden meer terrorisme gerelateerde tweets worden geplaatst, maar vermoeden dat hun methode ook omgekeerd kan worden gebruikt: afhankelijk van welke data je bekijkt kan een verhoging



Figuur 3.2: Terrorismegerelateerde tweets in de tijd.

in het aantal terrorisme gerelateerde tweets ook duiden op een aanstaande terroristische aanval (met name als deze activiteit plaatsvindt onder internetgebruikers van wie al bekend is dat ze geradicaliseerd zijn). Dergelijke informatie kan als waarschuwing dienen en kan daardoor zeer nuttig zijn voor veiligheidsinstanties.

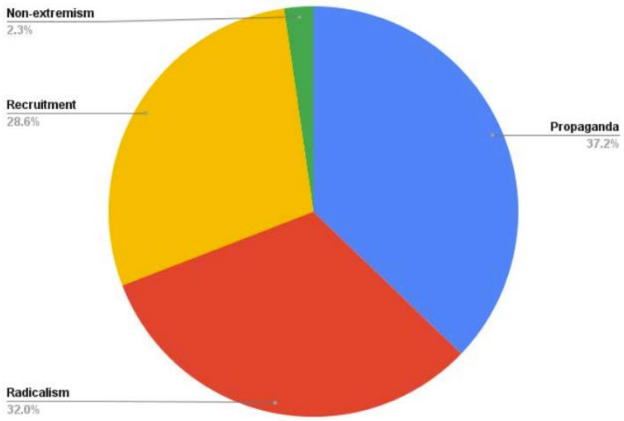
3.1.4 Het classificeren van radicalisering

Rajendran et al. (2022) hebben een methode ontwikkeld om specifiek te kijken naar de intentie van tweets afkomstig van een geradicaliseerd persoon. Ze maken hierbij onderscheid tussen een tweet die simpelweg blijk geeft van radicalisering, een tweet die propaganda bevat en een tweet die bedoeld is om mensen te rekruteren. De dataset die de onderzoekers gebruiken bestaat uit meer dan 93.000 tweets die geplaatst zijn in de aanloop naar en gerelateerd zijn aan de bestorming van het Capitool in de VS in 2021. Om deze tweets correct te classificeren gebruiken de auteurs twee soorten modellen: een zogenaamd

Long Short-Term Memory-netwerk (LSTM) en drie verschillende soorten *BERT (Bidirectional Encoder Representations from Transformers) modellen*.[16] Zie Bijlage A voor een beschrijving van deze modellen.

Rajendran et al. (2022) hebben onderzocht of deze modellen erin slaagden om te bepalen of een bepaalde tweet radicale content bevat en of het daarnaast ook bedoeld is om te rekruteren of propaganda te verspreiden. De LSTMs bleken dit goed te kunnen, met een hoogste F-score van 0,86, maar de BERT-modellen scoorden beter. Het best scorende model is een specifieke versie van BERT, *RoBERTa*, en haalde een F-score van 0.93.

Dit betekent dat deze modellen nauwkeurig kunnen voorspellen welk doel een tweet van een geradicaliseerd persoon dient. Zoals de distributie van de tweets uit de dataset van Rajendran et al. in Figuur 3.3 laat zien bestaat hun dataset vrijwel uitsluitend uit tweets die geschreven zijn door geradicaliseerde personen. Dit betekent dat hun model niet geschikt is om radicalisering op te sporen (daar zijn deze modellen niet op getraind, aangezien ze nauwelijks voorbeelden hebben gezien van tweets die *niet* van geradicaliseerde personen zijn) maar voor een mogelijke stap daarna: het fijnmaziger classificeren van tweets met radicale content. Dit kan nuttig zijn voor overheidsinstanties omdat bepaalde tweets van geradicaliseerden mogelijk gevaarlijker zijn dan anderen: een tweet waarin iemand ontevredenheid uit over de overheid is van een ander kaliber dan een tweet waarin iemand oproept een overheidsgebouw te bestormen.



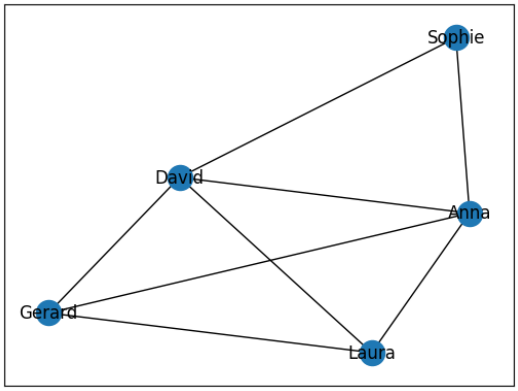
Figuur 3.3: Distributie van tweets in de dataset van Rajendran et al. (2022).

[16] Te weten: BERT, DistilBERT en RoBERTa.

3.2 Sociale netwerkanalyse

Naast NLP-technieken worden ook technieken uit het veld van Social Network Analysis (SNA) gebruikt voor het opsporen van geradicaliseerde gebruikers op sociale media. SNA richt zich op het in kaart brengen en analyseren van netwerken. Dit gaat bijvoorbeeld om netwerken tussen personen, computers of bedrijven.

In Figuur 3.4 staat een voorbeeld van een representatie van een sociaal netwerk.[17]



Figuur 3.4: Voorbeeld van een representatie van een sociaal network.

De cirkels representeren de entiteiten in het netwerk en de lijnen de verbindingen tussen de entiteiten (in dit geval personen). Een centraal concept binnen de SNA is *centraliteit*: een maatstaaf voor hoe groot iemands invloed is in het netwerk. De meest simpele manier om centraliteit te meten is door te kijken naar hoeveel verbindingen elke entiteit heeft met andere entiteiten in het netwerk: de zogenaamde *degree centrality*. In dit geval zijn David en Anna de entiteiten met de hoogste *degree centrality*: zij zijn elk verbonden met alle andere vier entiteiten in het netwerk. Sophie heeft de laagste *degree centrality*, met twee verbindingen met andere entiteiten. Een andere veelgebruikte methode is *betweenness centrality*. Een entiteit heeft een hoge *betweenness centrality* als ze zich vaak in het kortste pad tussen twee andere entiteiten bevindt. In Figuur 3.1 geldt dit voor David en Anna, die op het kortste pad tussen Gerard en Sophie en het kortste pad tussen Laura en Sophie zitten.

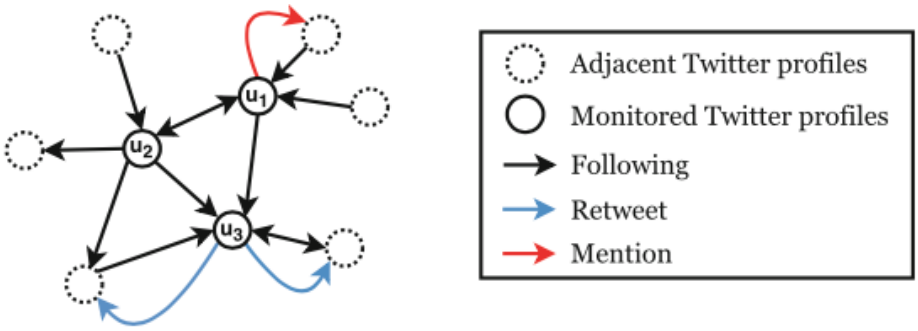
Deze techniek wordt op twee manieren gebruikt voor het detecteren van online extremisme:
1. het vinden van mogelijk geradicaliseerde personen;
2. het in kaart brengen van netwerken van geradicaliseerde personen.

In het eerste geval gaat het om het identificeren van mensen die mogelijk geradicaliseerd zijn. Dit gebeurt niet door het analyseren van de door hen geschreven berichten maar door het bekijken van hun positie in bepaalde netwerken: wanneer een gebruiker van een sociaal netwerk berichten uitwisselt of deelt van andere gebruikers van wie bekend is dat ze geradicaliseerd zijn, is de kans groter dat deze gebruiker zelf geradicaliseerd is.

[17] Dit voorbeeld is gegenereerd met het Python NetworkX-package, zie https://networkx.org/ .

3.2.1 Het opsporen van geradicaliseerde personen met netwerktechnieken

Een voorbeeld hiervan is het werk van López-Sánchez, Corchado, & González Arrieta (2018). Deze auteurs hebben een systeem ontwikkeld waarbij begonnen wordt met een aantal Twitteraccounts waarvan bekend is dat deze horen bij geradicaliseerde mensen. Accounts die gelinkt zijn aan deze geradicaliseerde accounts door middel van het volgen, re-tweeten of 'mentionen' van deze accounts worden in kaart gebracht zoals in Figuur 3.5.



Figuur 3.5: SNA-methode van López-Sánchez et al. (2018).

Hier zien we drie accounts (u_1, u_2 en u_3) van geradicaliseerde personen en hun relaties met andere, nog onbekende, accounts. De accounts die gelieerd zijn aan de geradicaliseerde accounts worden eerst met een simpele formule gerangschikt. Deze formule houdt rekening met de hoeveelheid interacties: hoe meer interactie met de geradicaliseerde account, hoe hoger de score. Daarnaast speelt het zogenaamde *sentiment* van tweets met mentions van een geradicaliseerde account een rol: als de gebruiker zich positief uitlaat in een tweet waarin de geradicaliseerde gebruiker wordt gementioned, verhoogt dit de score, en als het sentiment negatief is verlaagt dit juist de score.[18] De accounts met de hoogste scores worden aan de database toegevoegd en gecontroleerd door een mens. Andere voorbeelden van werk waarin SNA en soortgelijke methoden worden gebruikt om geradicaliseerde accounts op te sporen zijn Agarwal & Sureka (2015), Gialampoukidis et al. 2017, Petrovskiy & Chikunov 2019 en Sánchez-Rebollo et al. (2019).

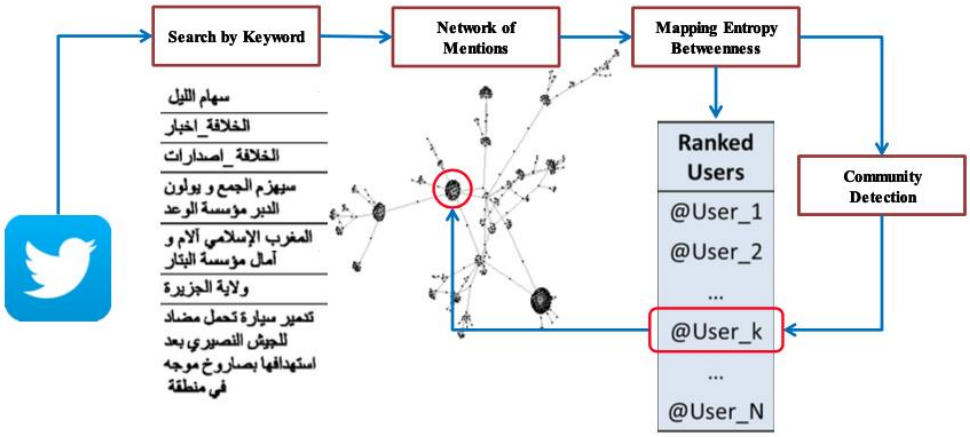
3.2.2 Het in kaart brengen van netwerken van geradicaliseerde personen

In het tweede geval is het doel niet alleen de identificatie van geradicaliseerde personen, maar ook het analyseren van de relaties tussen geradicaliseerde personen in een netwerk. Hierbij speelt centraliteit een belangrijke rol: mensen met een hogere centraliteitsscore hebben over het algemeen meer invloed in het netwerk dan mensen met een lagere score en kunnen dus belangrijke gebruikers zijn om in de gaten te houden.

Een voorbeeld van deze strategie is te vinden in Gialampoukidis, Kalpakis, Tsikrika, Vrochidis, & Kompatsiaris (2016), wier werk zich richt op het in kaart brengen van netwerken van jihadistische twittergebruikers. Zoals in Figuur 3.6 te zien is beginnen de auteurs met het verzamelen van relevante tweets op basis van een lijst Arabische trefwoorden die te maken hebben met terroristische propaganda. Op basis van de gevonden accounts die tweets

[18] Hierna wordt nog een analyse gedaan van de profielfoto's van de betreffende accounts.

hebben gepost met de relevante trefwoorden, wordt door middel van *mentions* een netwerk van potentiële extremistische accounts in kaart gebracht.



Figuur 3.6: Methodiek van Gialampoukidis et al. (2017).

De volgende stap is *Mapping Entropy Betweenness*, een maatstaaf die afgeleid is van de *degree centrality*. Bij deze stap worden de centrale spelers in het netwerk van de gevonden accounts geïdentificeerd. Vervolgens vindt *community detection* plaats: de accounts die contact hebben met de centrale spelers worden gemarkeerd als horend bij de *community* van de centrale spelers. Op deze manier wordt een complex netwerk samengesteld dat laat zien welke accounts belangrijk zijn en welke accounts zich om deze accounts heen bevinden. Andere onderzoeken naar deze methode zijn te vinden in bijvoorbeeld Bermingham, Conway, McInerney, O'Hare, & Smeaton (2009), Gialampoukidis et al. (2016), Benigni, Joseph, & Carley (2017) en Moussaoui, Zaghdoud, & Akaichi (2019).

3.3 Large Language Models

Sinds November 2022 – de lancering van de publieke versie van ChatGPT door OpenAI- zijn Large Language Modellen (LLMs, “grote taalmodellen”) niet meer weg te denken uit de AI. LLMs zijn zogeheten generatieve modellen, die, op basis van context, gebruikersinput en trainingsdata, nieuwe teksten genereren. Een taalmodel is *groot* als het getraind is op veel woorden, veel computerkracht heeft benut tijdens zijn training (GPU FLOPS), en veel *parameters* heeft: variabelen die hun performance bepalen en dus zorgvuldig moeten worden ingesteld. LLMs zijn gebaseerd op neurale netwerken (zogeheten Transformers, Vaswani *et al.*, 2017), en hun parameters bestaan uit de gewichten van de connecties tussen de verschillende neuronen in de onderliggende netwerken.

LLMs vormen een spectrum, met aan de linkerkant de zogeheten *base* of *foundation* models: basismodellen die geen instructies van gebruikers opvolgen, maar “slechts” goed zijn in het genereren van lopende tekst. Basismodellen die worden getraind op het opvolgen van menselijke instructies bevinden zich in het midden van het spectrum. Aan de hand van zogeheten *prompts* kunnen deze modellen gespecificeerd input-output gedrag vertonen, aan de hand van door de mens geprepareerde voorbeelden. Die voorbeelden kunnen worden aangeboden aan het vluchtige werkgeheugen van LLMS, of persistent gemaakt worden door de modellen bij te trainen op dit materiaal (“fine-tuning”).

Aan de rechterkant van het spectrum treffen we “AI-assistenten” aan als ChatGPT en Bard. Deze modellen zijn verder gefine-tuned op menselijke voorkeuren, via een kostbaar en

intensief proces dat “Reinforcement Learning from Human Feedback” (RLHF) heet. RLHF is van groot belang voor het verwezenlijken van communicatieve omgang met LLMs – via dialoogvoering kan de menselijke gebruiker “onderhandelen” over de aard en vorm van de gewenste informatie. Hierbij moet worden opgemerkt dat LLMs in het algemeen niet sterk zijn in het reproduceren van feitelijke informatie, en regelmatig onvoorspelbaar gedrag vertonen (“hallucineren”), zowel bij het presenteren van feiten als tijdens hun redeneringen. Dit heeft geleid tot technische aanpassingen waarmee LLMs kunnen communiceren met externe databases, voor het ophalen van gecureerde, actuele en betrouwbare informatie.

Alhoewel generatieve taalmodellen al decennia bestudeerd en toegepast worden, zijn neurale taalmodellen door hun kracht en schaal een doorbraak. Cruciaal ingrediënt is de conditionering van het woordgeneratieproces op interne informatie uit de onderliggende neurale netwerken (zoals vector-representaties van woorden in de context), in tegenstelling tot de letterlijke woordcontext in de standaard, niet-neurale taalmodellen. Hierdoor werken LLMs intern met zeer rijke, contextuele representaties van taal. LLMs kunnen evenwel meer dan alleen taal genereren. Zo kunnen LLMs, na prompting of fine-tuning, natuurlijke taalanalyses uitvoeren, en reflecteren op allerlei aspecten van taal. Een Engels voorbeeld met ChatGPT is:

- Gebruiker: What is extremist in this sentence: "People should not obey any government"
- ChatGPT: *In the sentence "People should not obey any government," the term "extremist" is not explicitly used. However, the statement itself can be considered an extreme or radical position, as it advocates for complete disobedience to any form of government authority. The term "extremist" generally refers to individuals or groups who hold extreme views or engage in actions outside the mainstream or accepted norms.*

Dit illustreert dat bepaalde LLMs, zoals de AI-assistent ChatGPT, potentieel inzetbaar zijn als detectietools, met het voordeel van bijbehorende leesbare uitleg. Nader onderzoek is nodig om hun geschiktheid systematisch te beoordelen, met als complicerende factor de dynamiek van het LLM-veld, de statistische aard van deze modellen (zodat herhaalde experimenten niet altijd dezelfde resultaten geven) en het feit dat veel benchmark data (waarmee machine learning modellen getest en vergeleken kunnen worden) op het web staat, waardoor de kans groot is dat veel LLMs dergelijke data al “gezien” hebben, en ervan geleerd hebben.

Er zijn ook minder goedaardige toepassingen van LLMs denkbaar. Door hun grote mate van beschikbaarheid en laagdrempelige toegang vormen LLMs interessant en nieuw gereedschap voor kwaadwilligen die goedkoop en snel toxische, extremistische online content willen genereren. Uit intern TNO onderzoek is gebleken dat het relatief eenvoudig is om een *bona fide* LLM toxische content te laten genereren, via een finetuningsproces dat slechts een uur in beslag neemt. Op dit moment circuleren er duizenden soortgelijke, potentieel “beïnvloedbare” open source LLMs (met name op hubs als Huggingface). Daarnaast kunnen LLMs worden ingezet om de “workflow” van online extremisten te ondersteunen – meertalige LLMs kunnen snel en geautomatiseerd vertalingen van extremistische content maken naar andere talen, bijvoorbeeld, of hun stijl en sentimenten variëren en aanpassen aan bepaalde groeperingen.

3.4 Conclusies

In deze subsectie hebben we op basis van literatuuronderzoek gekeken naar technieken uit de NLP-hoek en uit de SNA-hoek om automatisch te bepalen of een bepaalde tekst radicale content bevat en wat voor type radicale content dit is (zoals propaganda of rekrutering). Daarnaast zijn er technieken besproken die geradicaliseerde personen volgen in de tijd en

die nieuwe geradicaliseerde accounts opsporen aan de hand van al bekende geradicaliseerde accounts. De technieken besproken in deze publicatie zijn mogelijk interessant voor praktische toepassingen, maar de publicaties rapporteren alleen over eigen onderzoeksdata. Een volgende stap kan bestaan uit het implementeren en evalueren van de diverse aanpakken op eigen data. Een interessante aanvulling op content-gebaseerde detectie is het betrekken van gedragsmatige kenmerken van verspreiders en consumenten van online radicale of extremistische content, zoals sociale netwerk-informatie, post- en volg-gedrag in de digitale ruimte, en subtiele gedragsmatige indicatoren van vatbaarheid en weerbaarheid voor deze content.

3.5 Detectiemethodieken uit de praktijk

Het doel van detectie is om potentieel schadelijke activiteiten op te sporen voordat ze escaleren en om de autoriteiten of andere relevante instanties op de hoogte te stellen, zodat ze passende maatregelen kunnen nemen om de dreiging te verminderen of te voorkomen.

Er kan onderscheid gemaakt worden tussen de detectie vanuit overheidspartijen (e.g. NCTV, inlichtingendiensten) en rechtshandhavingspartijen (e.g. politie, Europol) en vanuit bedrijven (e.g. service providers, sociale media platforms, hosting providers). Alle werken met verschillende doelstellingen voor de detectie van content en hebben ook andere grondslagen om hier mee te werken.

Voor een beschrijving van detectiemethodieken in de praktijk hebben we gebruikt gemaakt van de interviews die zijn gehouden met de verschillende stakeholders. Hun informatie is aangevuld met literaire bronnen.

3.5.1 Detectie vanuit overheidspartijen

Detectie en duiding

Binnen overheidspartijen wordt vooral gedetecteerd op taalgebruik en het gebruik van extremistische afbeeldingen en symbolen. Waar de service providers zich richten op content moderatie, wordt de detectie door overheidspartijen ingezet met twee hoofddoeleinden. Ten eerste wordt detectie gebruikt om de trends op het gebied van radicalisering binnen de maatschappij in kaart te brengen (fenomeen niveau). Ten tweede is detectie een manier om vast te stellen of er meer onderzoek nodig is naar een specifieke casus (individu niveau) om te kijken of de plaatser van bepaalde content wellicht veiligheidsrisico's veroorzaakt.

Niet alle overheidspartijen mogen zelf content detecteren. De overheidspartijen die wel gebruik mogen maken van contentdetectie gebruiken automatische detectietools en een team van contextduiders, informatiemedewerkers die informatie duiden en beoordelen. Deze medewerkers besluiten of content daadwerkelijk schadelijk, potentieel gevaarlijk, gewelddadig of radicaliserend is.
De medewerkers richten zich dan op het verwerken van de content door het te verwijzen naar de juiste partij. Zo wordt er besloten of de content onderdeel is van een groter geheel, e.g. een netwerk of een nieuwe trend, en of de maker van de content een veelpleger is. Zij duiden dus de context waarin de ongewenste content plaats vindt en zijn de start van het proces waarin besloten wordt hoe er gehandeld dient te worden.

Technologiegebruik Nederlandse overheid

Binnen de Nederlandse overheid is er een reeks overheidsorganen die op een aantal manieren aan detectie doen. Technologische detectie wordt typisch ingezet bij inlichtingen- en opsporingsdiensten. Zo wordt sociale media monitoring software in opsporingsonderzoeken gebruikt of voor inlichtingsdoeleinden, waarbij men werkt vanuit verschillende grondslagen zoals de Wet op de Inlichtingen- en Veiligheidsdiensten (Wiv), of het inzetten van bijzondere opsporingsmiddelen zoals die zijn vastgelegd in de politiewet. Daarnaast is er bij veel overheidsdiensten sprake van *mediawatching* vanuit communicatiedoeleinden, *webcare* vanuit overheidsdienstverlening en zijn er nog specialistische toepassingen, zoals Real-Time Intelligence Centers bij de politie die meekijken op internet bij incidentopvolging in de meldkamer. Er zijn ook gemeenten die sociale media monitoring software gebruiken of gebruikten (Bantema, 2021).

Over het algemeen worden er commerciële sociale media monitoring tools gebruikt voor het detecteren van content op de grote sociale media platformen, hoewel dit gebruik de laatste jaren sterk is teruggebracht. Enerzijds omdat de platformen dit niet meer mogelijk maken voor overheidsdoeleinden of niet willen dat veiligheidsdiensten meekijken, anderzijds door nieuwe wetgeving als de AVG. In sommige gevallen, zoals voor specialistische analyses in opsporing of inlichtingen, worden speciale open-source tools gebruikt of eigen maatwerktools gemaakt.

Naast het gebruik van specialistische sociale mediamonitoringtools worden ook zoekmachines en sociale media zelf ingezet. Een voorbeeld van dit laatste zijn digitale wijkagenten die in hun gebiedsgebonden werk op diverse sociale mediaplatformen actief zijn. Zij kunnen in contact staan met hun volgers en meldingen krijgen, maar in het kader van de openbare orde ook zelf een digitale surveillance op een platform naar keuze doen. Ook daarin is het zoeken door tijdslijnen, zoeken met keywords en andere zoekmogelijkheden die de platforms toestaan mogelijk.

Het gebruik van artificial intelligence tenslotte wordt in een aantal deelprocessen van veiligheidsdiensten ingezet. Zo wordt machine learning-technologie ingezet om door grote hoeveelheden data te zoeken naar specifieke gegevens of in de hoeveelheid data patronen te vinden.

Kennisuitwisseling

Er wordt door overheidsinstanties gebruik gemaakt van kennisdatabases die gedeeld worden met andere stakeholders in hun werkveld. Een voorbeeld is de symbolenbank van de NCTV. Ook wordt subversieve content die herhaaldelijk wordt gedeeld op online platforms opgenomen in een hash database. Hash databases maken gebruik van *hashfuncties* om content te koppelen aan unieke digitale vingerafdrukken (“hashes”). Afbeeldingen worden op deze manier een waarde toegewezen die kan worden opgeslagen in een database, waardoor het snel opsporen van overeenkomsten en vergelijkbare afbeeldingen mogelijk wordt gemaakt.

De tools die worden gebruikt door de overheid worden niet uitgewisseld met service providers, omdat de kloof tussen private en publieke sector hier nog te groot is. De belangen van de verschillende partijen zijn zo anders dat het evenmin vanzelfsprekend is om alle informatie te delen. Zo is de grondslag voor overheidspartijen vaak gericht op het bewaken van de nationale veiligheid en het monitoren van relevante fenomenen. De private sector daarentegen heeft financiële baat bij het aanbieden van commerciële services, met gevolgen voor de manier waarop er met data wordt omgegaan; verzamelde gebruikersdata

heeft immers commerciële en bedrijfseconomische waarde, en wordt doorgaans niet zonder meer gedeeld. Verder is het juridische raamwerk waarbinnen publieke en private partijen werken te verschillend om vrije kennisdeling te garanderen.

Juridisch raamwerk

Binnen overheidspartijen zit er een groot verschil in de grondslag en het mandaat dat zij hebben om zelf content te mogen detecteren. Er kan globaal een onderscheid gemaakt worden tussen besturende organisaties en uitvoerende organisaties zoals in de rechtshandhaving. Inlichtingendiensten mogen in het kader van de Wiv zelf geen content detecteren. De uitvoerende organisaties in de rechtshandhaving vinden hun grondslag(en) in andere wetgeving (zoals de politiewet). De recent geïntroduceerde Autoriteit Terroristisch en Kinderpornografisch Materiaal (ATKM) valt onder het bestuursrecht. De verschillende wetgevingen zorgen er voor dat alle partijen verschillende grondslagen hebben voor het detecteren van en interveniëren bij online extremistische content.

Overheidspartijen zijn gehouden aan het juridische raamwerk waarin zij opereren. Het is belangrijk om als overheid een mate van onafhankelijkheid te bewaren en de privacyrechten van burgers te respecteren. Deze situatie creëert een knelpunt: het is ook belangrijk om te voorkomen dat extremistische en radicaal gedachtegoed zich verspreidt, en escaleert naar geweld.

Om de Wiv te honoreren wordt er in alle detectieprocessen binnen de overheidsdiensten gezorgd voor de ‘human in the loop’: een menselijke analist die de zoektermen en bronnen waarin gezocht wordt kan afbakenen om niet buiten de mandaten voor de detectie te treden. Verder worden de resultaten van de detectie onderworpen aan menselijke beoordeling. Uiteraard bestaat er hierbij een zorgvuldig proces waarin juridische en ethische afwegingen en toetsing op aanbesteding een rol spelen voordat tools geselecteerd en gebruikt worden.

Doelstellingen

Er worden geen *Key Performance Indicators* ingezet door de overheidsinstellingen. Hoewel er wel wordt bijgehouden hoeveel content er wordt gedetecteerd, worden er geen concrete kwantitatieve en kwalitatieve doelen aan verbonden.

3.5.2 Detectie vanuit commerciële bedrijven

Detectie en duiding

Elke serviceprovider of online platformaanbieder heeft eigen voorwaarden en eigen beleid rondom online gedrag. Wanneer aangemerkte content of een account zich hier niet aan houdt heeft de provider een grondslag om deze te verwijderen. Dit gebeurt in enkele gevallen automatisch, wanneer de content duidelijk grensoverschrijdend is en niet strookt met de beleidsregels. Hiervoor wordt veel gebruikt gemaakt van automatische detectie.

Naast detectie gericht op tekst, afbeeldingen en symboliek kan extremistische content ook indirect gedetecteerd worden via *metadata*: data “over data”, zoals temporele patronen, hoeveelheden data, en post- en deelgedrag. Bepaalde ongewenste content wordt geplaatst platforms met *end-to-end encryption*[19] In dergelijke gevallen wordt er gebruik gemaakt van

[19] End-to-end-encryptie verwijst naar een beveiligingsmethode waarbij communicatie, zoals berichten, alleen door de afzender en de ontvanger kan worden gelezen. Tijdens verzending is de inhoud versleuteld en alleen

analyse van metadata, zoals de frequentie van berichten of het gedrag van bepaalde gebruikers. Analyse van dit soort metadata kan tot op zekere hoogte indicaties geven van de aard en samenstelling van de onderliggende object data zonder die data te inspecteren. Serviceproviders zijn voor daadwerkelijke individuele opvolging sterk afhankelijk van informatie door anderen, (gebruikers of autoriteiten).

Er is weinig zicht op detectie van content bij kleinere platforms. Daar worden twee redenen voor gegeven. Ten eerste, de kleinere platforms hebben vaak niet de capaciteit om zich te richten op uitgebreide content moderatie. Zij kunnen hierin ondersteund worden door grotere platforms en initiatieven zoals Tech Against Terrorism en GIFCT. Beide richten zich op kennisdeling en bieden hun oplossingen en open-source middelen aan. Ten tweede, kleinere platforms hebben niet altijd de motivatie om deze content aan te pakken. Dit komt veelal door het verdienmodel dat er achter zit. Hier is een rol voor de overheid weggelegd, om middels wetgeving deze platforms toch te stimuleren zich te richten op content detectie en moderatie.

Technologiegebruik

Bij de detectie van taal en symbolen wordt in toenemende mate gebruik gemaakt *van Large Language Models* (zie subsectie 3.1) en zogeheten *scraping tools* die content extraheren uit websites en andere online bronnen. Gezamenlijk worden deze instrumenten ingezet om het proces van menselijke detectie te automatiseren. AI speelt hierbij een rol bij het aanleren van menselijk inzicht aan machine learning-modellen. Binnen de door ons geraadpleegde platformen wordt er verschillend beleid van contentmoderatie gevoerd. Enkele platformen kiezen er voor om content zowel automatisch te detecteren als te verwijderen. Als gebruiker is het altijd mogelijk om hier bezwaar tegen te maken en wanneer het bezwaar wordt toegekend, wordt de content weer terug geplaatst. Content wordt dus eerst verwijderd en daarna, bij beklag, gecontroleerd door mensen die de context kunnen duiden. Andere platformen kiezen er voor om content pas na menselijke (handmatige) duiding te verwijderen.

Kennisuitwisseling

Op het moment zijn Tech Against Terrorism en GIFCT de instanties die het meest genoemd worden als partners. Deze onafhankelijke bedrijven werken samen met de verschillende sectoren binnen de onderwerpen terrorisme, extremisme en radicalisering, waaronder overheden, service providers en social media platformen. Door een neutrale houding aan te nemen en te focussen op het creëren van vertrouwen tussen de verschillende partijen versterken zij de samenwerking tussen commerciële partijen met overheden. Daarnaast vervullen zij een rol in het ondersteunen van kleinere platforms in het tegengaan van ongewenst materiaal.

Er wordt door verschillende platformen gewerkt aan een verbeterde samenwerking. Zo is Altitude een voorbeeld van een samenwerking tussen Jigsaw (Google), Tech Against Terrorism en GIFCT. Het doel van deze gratis en *open source interventie* is om kleine en medium platforms in staat te stellen om duidelijke en onderbouwde beslissingen te maken op hun eigen sites wat betreft content moderatie. Zij kunnen zo een beleid maken dat zich richt op het bestrijden van extremistische content en hebben zij ook de middelen om hier op te handhaven.

ontcijferbaar door de beoogde ontvanger, waardoor derden, zelfs dienstverleners, geen toegang hebben tot de onversleutelde informatie.

Juridisch raamwerk

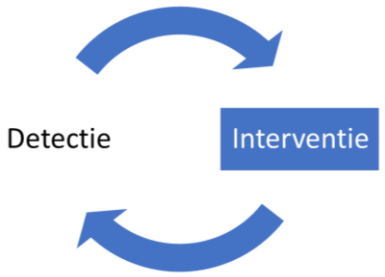
Ook de service providers hebben te maken met een juridisch raamwerk. Als knelpunt wordt benoemd dat er nog weinig uniformiteit is wat betreft de wetgeving rondom dit onderwerp in verschillende landen. Er worden daarbij ook verschillende definities gehanteerd, wat het lastig maakt om per land te voldoen aan de juiste eisen. Dit kan er voor zorgen dat kleinere platforms beperkt blijven tot bepaalde regio's of nooit aan eisen zullen kunnen voldoen en illegaal opereren binnen bepaalde landen.

Doelstellingen

Er worden momenteel geen *Key Performance Indicators* ingezet door de service providers. Hoewel er wel wordt bijgehouden hoeveel content er wordt gemodereerd, worden er geen concrete (kwalitatieve en/of kwantitatieve) doelstellingen aan verbonden.

4 Interventies bij online radicalisering en extremisme

Op basis van detectie van desinformatie of tekenen van online extremisme kunnen interventies worden ingezet. In dit hoofdstuk wordt een selectie van mogelijke of daadwerkelijke fysieke en online interventies beschreven die online verspreiding van extremistische informatie moeten tegengaan. Dit overzicht is voor een deel gebaseerd op wetenschappelijk onderzoek naar het voorkomen of tegengaan van desinformatie, en voor een deel op gesprekken met stakeholders vanuit overheid en commerciële instellingen.



4.1 Soorten interventies

In de context van online radicalisering verwijzen interventies naar preventieve of repressieve acties en maatregelen die worden genomen om te voorkomen dat informatie verder verspreid wordt en individuen (verder) radicaliseren of om mensen die al geradicaliseerd zijn te deradicaliseren en (verdere) schade bij mensen of in de omgeving te voorkomen of beperken. Deze interventies zijn gericht op het verminderen van de invloed van extremistische ideologieën en het begeleiden van mensen naar een minder radicale koers.

Benadrukt moet worden dat veel van de literatuur betrekking heeft op Westerse radicalisering of radicalisering in niet-Westerse landen die zich richt op het Westen als de vijand. De aannames die gedaan worden zullen dus niet overal ter wereld van toepassing zijn. Zo is bijvoorbeeld de mate waarin sociale normen of uitingen van een autoriteit het gedrag van mensen beïnvloeden verschillend over culturen heen (Cialdini, 2001).

Zoals eerder vermeld kan er onderscheid gemaakt worden tussen proactieve en reactieve interventies. Proactieve interventies tegen online radicalisering richten zich op het voorkomen van radicalisering en extremisme vóór de daadwerkelijke gebeurtenis. Deze benadering concentreert zich op het verminderen van risicofactoren, het bevorderen van veerkracht en het ontmoedigen van potentiële rekruten van extremistische ideologieën.

Aan de andere kant zijn reactieve interventies gericht op het omgaan met reeds opgetreden informatieverspreiding die mogelijk tot radicalisering kan leiden. Ze zijn gericht op individuen of groepen die al geradicaliseerd zijn en omvatten strategieën zoals wetshandhaving en maatregelen om de directe dreiging die van hen uitgaat te verminderen. Deze aanpak is specifieker en richt zich op het aanpakken van bestaande extremistische gedragingen. In de praktijk worden vaak zowel proactieve als reactieve strategieën gecombineerd om een alomvattende aanpak te bieden bij het bestrijden van online radicalisering en extremisme. Tabel 4.1 geeft een overzicht.

Tabel 4.1: Voorbeelden van interventiemogelijkheden.

	Fysiek (offline)	Online
Proactief	Bevorderen veerkracht van personen of groeperingen, benaderen rekruten en vatbare individuen, infiltratie binnen fysieke (non-digitale) netwerken.	Mengen in online berichtenfora, verspreiden van "counter-narratives" en online beïnvloeding van meningen.
Reactief	Wetshandhaving, juridische maatregelen	Deactiveren van accounts, content of netwerken.

4.2 Proactieve interventies

We kunnen proactieve interventies onderscheiden in het fysieke en in het online domein. Veel relevante interventies zijn te vinden in literatuur over desinformatie en misinformatie. Deze interventies zijn allemaal gericht op de eerste, vroege stadia van radicalisering.

Proactieve interventies kunnen in verschillende stadia van het radicaliseringsproces worden ingezet, met een focus op verschillende doelgroepen en thema's (Van Wonderen, et al., 2023). Zo kunnen interventies zich richten op het brede publiek, met als doel weerbaarheid tegen extremisme te vergroten. Daarnaast kunnen proactieve interventies zich richten op personen die al tekenen van radicalisering laten zien, of op de rol van die personen in een netwerk. (Van Wonderen, et al., 2023) noemen in dit kader interventies die sleutelfiguren helpen beter problemen binnen de eigen gemeenschap te signaleren en te hanteren (Cankor & Noor, 2021). Naast het vergroten van de weerbaarheid van het individu en het actief betrekken van gemeenschappen blijken vooral counter-narratives relevant (Stephens, Sieckelinck & Boutellier (2021).

De focus van proactieve interventies tegen online radicalisering ligt op het moment van deze studie voornamelijk op educatie. Door de mediawijsheid van jongeren te vergroten worden ze minder vatbaar voor mogelijke online werving. Het vergroten van mediawijsheid is op het moment voornamelijk gericht op het aanpakken van mis- en desinformatie.[20]

Omdat het doel van bepaalde content niet altijd nauwkeurig te bepalen is, is het lastig om precies te duiden wat mis- en wat desinformatie is. In de context van radicalisering is desinformatie het intentioneel verspreiden van valse of misleidende informatie met als doel om personen of groepen extremistisch gedachtegoed te laten adopteren. De misinformatie die hiervoor gebruikt wordt is vaak in de vorm van bepaalde narratieven die ingezet worden om bepaalde politieke, religieuze of sociale ideologieën te promoten. Deze narratieven

[20] Misinformatie verwijst naar onjuiste informatie, die zowel opzettelijk als per ongeluk kan worden verspreid, terwijl desinformatie specifiek doelt op opzettelijk verspreide onjuiste informatie.

verkleinen complexe fenomenen uit de echte wereld tot simpele, geweld-verheerlijkende propaganda, hetgeen leidt tot radicalisering bij de consumenten van deze content (Sarah L. Carthy, 2023). Daarbij is ook uit onderzoek gebleken dat bepaalde karakteristieken van misinformatie goed resoneren bij personen die het risico lopen om te radicaliseren, met als risicofactoren sociale uitsluiting en lage cognitieve flexibiliteit (McCann, 2023).

4.2.1 Fysieke proactieve interventies

Een fysieke (ofwel offline) proactieve interventie is het beschikbaar stellen van adviezen om mediawijsheid te bevorderen. Dit soort adviezen geeft mensen strategieën om foutieve en misleidende informatie te herkennen. Social mediaplatforms zoals Facebook hebben dit soort adviezen geïmplementeerd; bijvoorbeeld "be sceptical of headlines," "look closely at the URL," and "investigate the source." (Kozyreva, et al., 2023). Een andere proactieve fysieke interventie is het expliciet maken van sociale normen. Dit kan de overheid of een social mediaplatform doen door te benadrukken dat de meeste mensen van een bepaalde groep het delen of gebruiken van foutieve informatie afkeuren en/of dat dergelijke acties over het algemeen als verkeerd, ongepast of schadelijk worden beschouwd (Kozyreva, et al., 2023).

4.2.2 Online proactieve interventies

Deze twee vormen van proactieve interventies zijn ook in online vorm te vinden. Daarnaast identificeren (Kozyreva, et al., 2023) een aantal zogenaamde *nudges*, letterlijk duwtjes: veranderingen in de omgeving van mensen die gebruik maken van het principe van snelle, automatische beslissingen en heuristieken (Kahneman, 2012). Een voorbeeld van een nudge is mensen vragen om te pauzeren en na te denken voordat ze inhoud op sociale media delen. Dit kan zo simpel zijn als een korte prompt, bijvoorbeeld: 'Wil je dit lezen voordat je het deelt?' (Kozyreva, et al., 2023). Deze techniek wordt 'friction' genoemd, en is een van de manieren waarop online nudges worden ingezet. Andere manieren zijn het saillant maken van sociale normen (zie ook onder fysieke interventies), het gebruik van labels om betrouwbaarheid van bronnen aan te geven, labels die waarschuwingen en factchecking weergeven, en zogenaamde accuracy prompts. Deze nudges zijn experimenteel onderzocht en er lijkt flink wat evidentie te bestaan voor het (in ieder geval op korte termijn) werken van deze nudges (zie bijvoorbeeld (Ecker, et al., 2022) (Epstein, et al., 2021) (Pennycook, McPhetres, Zhang, Lu, & Rand, 2020).

Naast nudges draait een aantal online proactieve interventies om het weerleggen van onjuiste informatie (*refutation*). *Debunking* is een voorbeeld van een interventie die gebruik maakt van het weerleggen van onjuiste informatie. Bij debunking worden valse beweringen blootgelegd (zie bij voorbeeld (Ecker, Butler, & Hamby, 2020) en (Schmid & Betsch, 2019). Een andere manier om in te gaan tegen onjuiste informatie is *counterspeech*. De Global Internet Forum to Counter Terrorism (www.gifct.org) definieert counterspeech als ''any online initiative, campaign, or engagement meant to promote positive alternatives or counter narratives to counteract the possible interest in terrorist and violent extremist groups''. Waar debunking uit de hoek van desinformatieliteratuur komt, richt counterspeech (ofwel counter-narratives) zich meer op het sociale aspect van argumenten. Counterspeech verwijst naar het gebruik van vrije meningsuiting en communicatie om te reageren op ongewenste of schadelijke ideeën, overtuigingen of retoriek. Het doel van counterspeech is vaak om tegenwicht te bieden aan extremistische, haatdragende, of anderszins schadelijke boodschappen door alternatieve, constructieve of verlichtende informatie te verspreiden. Verschillende onderzoeken hebben erop gewezen dat counterspeech effectief is in het verminderen van racistische haatspraak (Hangarter, Gennaro, Alasiri, & al., 2021)

(Obermaier, Schmuck, & Saleem, 2021) (Buerger, 2022), maar dit is wel afhankelijk van factoren zoals de toon van de counterspeech (Baider, 2023).

Uit onderzoek bleek dat de aanwezigheid en actieve steun van 'online gelijken' mensen aanmoedigden om zich actief in te zetten tegen haatspraak en de slachtoffers van dergelijke uitingen te ondersteunen (Garland, Ghazi-Zahedi, Young, Hébert-Dufresne, & Galesic, 2022). Wat van toepassing is 'in het echte leven' is ook van toepassing online: mensen voelen zich sterker samen en durven eerder in te grijpen bij bepaald gedrag wanneer zij weten dat zij gesteund worden door anderen.

Counterspeech biedt ook toepassingen voor het tegen gaan van online radicalisering. Het aanmoedigen van content die bepaalde radicaliserende of extremistische content tegenspreekt of zelfs afstraft, zou er wellicht toe kunnen leiden dat het draagvlak voor dit soort content verkleint onder de gebruikers van sociale media platformen. Counterspeech kan echter ook leiden tot een beweging naar kleinere platformen. Zo bleek uit een studie van dat na counterspeech en *counter radicalisation*-inspanningen bij sympathisanten van IS op Twitter, de groep zichzelf begon te censureren maar ook volgers aanmoedigde om zich te verplaatsen naar Telegram, waar ze niet publiekelijk zichtbaar waren (Mitts, 2021). Het publieke draagvlak voor extremistische uitingen werd dus kleiner, maar door de *counterspeech* verdween het fenomeen niet volledig.

Een ander voorbeeld van een proactieve interventie die gebruik maakt van het weerleggen van foutieve informatie is het concept van *inoculatie*. Vergelijkbaar met hoe een vaccinatie werkt, veronderstelt inoculatie binnen de context van desinformatie dat mensen een resistentie kunnen opbouwen voor misleidende en onware berichten door ze te confronteren met een verzwakte versie hiervan. Op deze manier zullen mensen sneller immuun zijn voor desinformatie als ze hiermee in het echte leven geconfronteerd worden (Van Wonderen & Peeters-Osseyran, Werken aan weerbaarheid tegen desinformatie en een eenzijdige meningsvorming, 2022). Uit verschillende onderzoeken is gebleken dat deze methode met name effectief is omdat deze de mediawijsheid van de gebruiker verhoogt. Uit onderzoek van Braddock bleek bijvoorbeeld dat gedragsinoculatie leidde tot weerbaarheid tegen beïnvloedingstechnieken die gebruikt worden in extremistische propaganda (Braddock, 2022). Daarom zou het nuttig kunnen zijn bij antiradicaliseringsinitiatieven. Een praktisch voorbeeld van de inoculatietheorie is *Radicalise*, een *serious game* waarbij de speler in de rol van een *recruiter* kruipt om een doelwit te radicaliseren.[21] Zo leert de speler de zwakke punten van een doelwit kennen en kan de speler daarna beter een radicaliseringsproces herkennen in zijn of haar omgeving. Wetenschappelijk onderzoek heeft aangetoond dat dit spel (en vergelijkbare spellen) effectief zijn in het weerbaar maken van personen tegen gevaarlijke invloeden (Saleh, Roozenbeek, Makki, McClanahan, & Van der Linden, 2021) (Roozenbeek, Van der Linden, Goldberg, Rathje, & Lewandowsky, 2022).

4.3 Reactieve interventies

Ook bij reactieve interventies kunnen we onderscheid maken tussen fysieke en online interventies. Sommige interventies hebben zowel een fysieke als een online vorm (zoals het opsporen van verspreiders van desinformatie), andere horen duidelijk in één van de domeinen thuis. Anders dan bij proactieve interventies zijn er in de literatuur nauwelijks geëvalueerde interventies te vinden. In dit overzicht putten we daarom voornamelijk uit interviews en workshops met stakeholders vanuit de overheid en commerciële instellingen,

[21] Radicalise: gratis lespakket bij serious game - Ter-info

en uit een overzicht “Toolbox of interventions against online misinformation V2.0” Kozyreva, et al., 2023).

4.3.1 Fysieke reactieve interventies

De meest voorkomende reactieve interventie is content moderatie naar aanleiding van detectie van de eigen systemen of mensen, of meldingen vanuit de buitenwereld. De content moderatie kan enerzijds worden toegepast door het verwijderen van ongewenste content, anderzijds kan het gaan om het blokkeren van bepaalde netwerken en accounts.

Veiligheidsdiensten hebben mandaat om in de ‘fysieke wereld’ in te grijpen wanneer er online bepaald gedrag plaats vindt. Zo kunnen er zogeheten STOP-gesprekken worden gevoerd, waarbij een persoon wordt aangesproken op hun gedrag online. Deze interventies worden casusgericht ingezet, in het bijzonder bij recidivisten. Mocht het gedrag online de wet overtreden is het ook mogelijk om personen te traceren en deze te vervolgen.

Bij commerciële bedrijven verschilt het huidige beleid in interventies per stakeholder. In het geval van serviceproviders sluit het beleid qua interventies vaak aan op het beleid van de detectiemogelijkheden. Het mandaat om te interveniëren bij bepaalde content stopt bij het verwijderen van content of accounts.

4.3.1 Online reactieve interventies

Waar overheidspartijen zich doorgaans beperken tot online proactieve interventies (zoals nudges) of fysieke reactieve interventies (zoals het blokkeren van accounts), zetten commerciële instellingen, zoals social media platforms, juist in op online reactieve interventies. In deze paragraaf worden voorbeelden van deze interventies besproken.

Deranking

Verschillende platformen maken gebruik van algoritmen om bepaalde content meer naar voren te plaatsen, zodat de gebruiker deze content eerder ziet. Een aantal interventiemethoden haakt in op die algoritmen om extremistische content tegen te gaan. Zo degraderen platformen bepaalde content (bijvoorbeeld extremistische teksten) in de feed van gebruikers of in de zoekresultaten; dit wordt *deranking* genoemd. YouTube zet deze methode in om bepaalde extremistische video’s te ontmoedigen. Dit resulteert overigens ook in een lagere omzet voor de plaatser van de content. Er zijn geen officiële bronnen van YouTube waarin het concept of proces van deranking wordt uitgelegd, maar er zijn veel blogs en posts door gebruikers op verschillende fora, waarin opgemerkt wordt dat bepaalde content niet organisch naar boven komt of zelfs actief niet zichtbaar wordt gemaakt.[22] [23] [24]

Verwijderen van content en accounts

Een andere manier om het verspreiden van extremistische content tegen te gaan is het verwijderen van die content of van hele accounts. Op het moment gebeurt dat bij verschillende platformen op verschillende manieren. De meest voorkomende is dat er door middel van een geautomatiseerd proces content wordt gemarkeerd. Deze content wordt

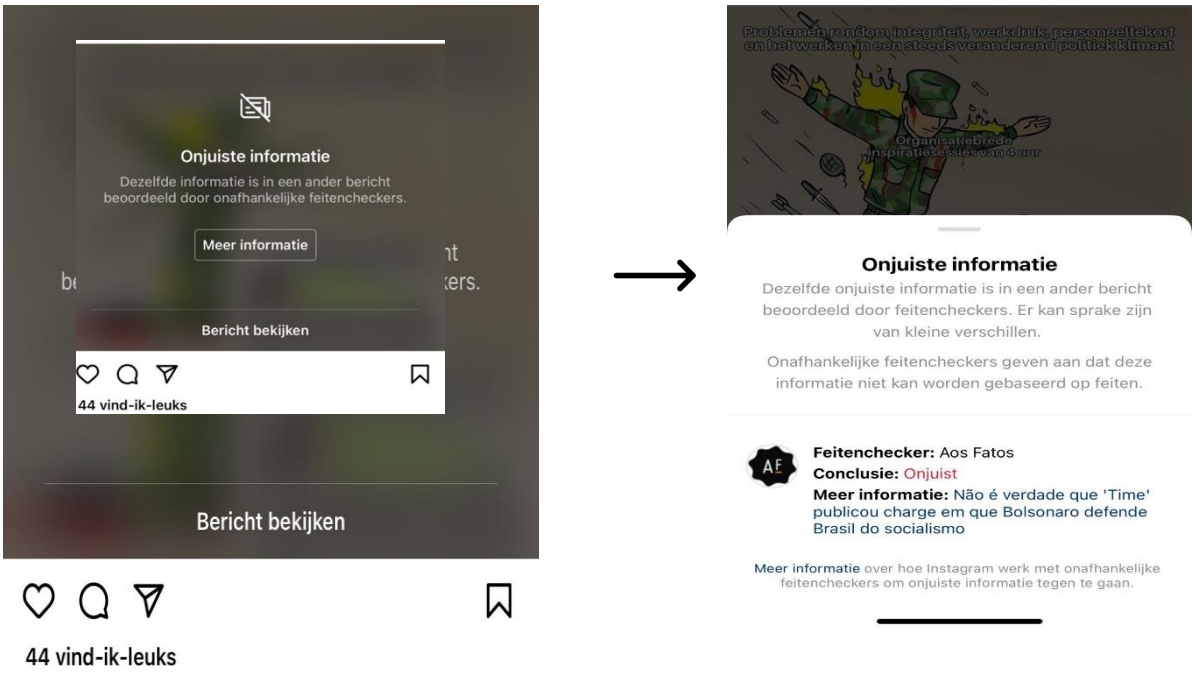
[22] [Misc FD] YouTube is artificially deranking conservative channels in search results : r/youtube (reddit.com)
[23] https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwiOr_7urZ6DAxWa1QIHHcrODxcQwqsBegQIDRAF&url=https%3A%2F%2Fwww.youtube.com%2Fwatch%3Fv%3DkOSorO-iBRg&usg=AOvVaw1rxhx0ejly4RZ0CTkst3HM&opi=89978449
[24] https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwj_koXcrp6DAxXEhqQKHZ5HBPU4ChAWegQICRAB&url=https%3A%2F%2Fwww.buzzfeednews.com%2Farticle%2Fellievhall%2Fyoutube-anti-meghan-markle-search-channels&usg=AOvVaw0jREGxDyEMc-YgCEb23JlE&opi=89978449

voorgelegd aan een menselijke medewerker, die dan inschat of de content daadwerkelijk tegen de richtlijnen ingaat en verwijderd dient te worden. Slechts één serviceprovider gaf aan bepaalde content automatisch te verwijderen.

Het is lastig om te bepalen hoe effectief deze interventie is. Sporadisch onderzoek laat zien dat het verwijderen van accounts van belangrijke individuen de consumptie en productie van haatdragende inhoud verminderen (Thomas & Wahedi, 2023). Bij de eerder genoemde verwijdering van de Boogaloo-beweging bleek na 18 maanden dat de kwantiteit van Boogaloo content op nagenoeg hetzelfde niveau was als voor de ban, alleen was de content zodanig aangepast (door het gebruik van verwijzingen) dat het niet meer direct aan te merken valt als Boogaloo content (Jigsaw, 2023).

Factchecken en markeren

Door bepaalde content te markeren als twijfelachtig of mogelijk onwaar wordt er van de gebruiker gevraagd kritisch te zijn over het al dan niet delen van deze content. Ter illustratie staat in Figuur 4.1 een voorbeeld van een Instagram post die door Instagram zelf is aangemerkt als mogelijk onjuiste informatie. De post zelf wordt gecensureerd en de gebruiker moet actief de keuze maken om de informatie toch te zien. Wanneer er wordt doorgeklikt krijgt de gebruiker toch de post te zien, maar met een waarschuwing dat de content mogelijk onjuist is. Wanneer er dan weer wordt doorgeklikt krijgt de gebruiker te zien waarom de post als zodanig gemarkeerd is (Figuur 11). Het nadeel van deze methode is dat er wordt uitgegaan van andere sociale media gebruikers om de content te markeren. Het kan dus zijn dat door vrijwilligers gemarkeerde content niet objectief onwaar is.



Figuur 4.1: Een post op Instagram die door Instagram zelf is aangemerkt als mogelijk onjuiste informatie.

Figuur 4.2: Geblokkeerde post op Instagram, post met waarschuwing, en uitgebreide informatie over de feitelijke correctheid van de post.

Verder heeft onderzoek uitgewezen dat het toevoegen van prompts, zoals 'do you want to read this before sharing', effectief is in het tegengaan van verspreiding van misleidende content (Bates, 2017). Dit soort prompts en labels zorgt er voor dat mensen zich bewuster zijn van de content die ze consumeren en verspreiden.

Bij het inzetten van markeren van bepaalde content als interventiemethode moet er samengewerkt worden tussen service providers en (overheids)experts, die kunnen duiden wat voor soort content ongewenst is en hoe die het beste geduid kan worden.

Eén van de geïnterviewde platformen heeft een functie toegepast waarmee het niet mogelijk is om naar een externe site te linken. Dat betekent dat bij bepaalde radicaliserende content er niet door verwezen kan worden naar een volgend platform. Dit heeft echter als nadeel dat het lastig is om bepaalde informatie te factchecken, omdat er niet doorgelinkt kan worden naar de bron.

Moderatie

Sommige platforms, zoals Discord, gebruiken moderatie, waarbij ook gebruik wordt gemaakt van vrijwilligers. Zo kunnen er binnen het platform groepen ontstaan die gecontroleerd worden door gebruikers, waarbij er een beroep wordt gedaan op hun morele verantwoordelijkheid. Gebruikers van Discord kunnen een eigen server modereren en krijgen vanuit het platform hulp om dit veilig te doen.[25] Moderators moeten ook een examen afleggen om te bewijzen dat zij bekend zijn met de regels en gebruiken van Discord.

Tegengaan van verdienmodellen

Georganiseerde extremistische groepen en extremistische *content creators* financieren zichzelf door onder andere donaties, lidmaatschapsgelden, het verkopen van merchandise, en winst uit reclame. Uit onderzoek is gebleken dat op meer marginale (*fringe*) *social media* sites 23% van de prominente accounts om financiële steun van hun volgers vragen (Ghosh

[25] Community Safety and Moderation | Discord

& Stocking, 2023). Andrew Tate, een voormalig professioneel kickbokser die nu voornamelijk bekend is als social media persoonlijkheid met extreemrechtse, misogyne en conservatieve opvattingen, vraagt bijvoorbeeld 8000 dollar om lid te worden van zijn 'war room'.[26] Dit netwerk, volgens Tate bedoeld om een gemeenschap te vormen die zich richt op het bevorderen van persoonlijke groei en succes, wordt momenteel onderzocht vanwege beschuldigingen van uitbuiting van vrouwen (Elshout, 2023). Naast de 'war room' heeft Tate ook een 'Hustler University' programma, waar je voor 50 dollar per maand leert hoe je geld kunt verdienen met een eigen bedrijf. De combinatie van het 'war room' en het 'Hustler University' abonnement leveren Tate en zijn broer 5 miljoen dollar per maand op (Murray, 2023).

Verdienmodellen en winstbejag spelen dus ook een rol in het creëren van extremistische content. Voor bepaalde *content creators* is het zelfs de enige bron van inkomsten, omdat zij door hun extremistisch gedachtegoed geen (voltijd) baan hebben (Leidig E. , 2021). Het opwerpen van belemmeringen voor deze verdienmodellen kan bijdragen aan een vermindering van extremistische content. Zo gebruikt YouTube *demonetization* (ontwaarden) om bepaalde content van het platform te weren.[27] YouTube hanteert een demonetiseringsbeleid om te bepalen welke video's in aanmerking komen voor advertentie-inkomsten en welke niet. Het beleid is ontworpen om ervoor te zorgen dat advertenties worden weergegeven bij content die voldoet aan bepaalde richtlijnen en normen. Hoewel hierdoor content vol haatspraak offline werd gehaald, leidde het ook tot veel klachten van *content creators* omdat het algoritme willekeurig leek (Alexander, 2018).

4.4 Conclusie

In dit hoofdstuk hebben we verschillende soorten fysieke en online interventies besproken die radicalisering of extreemrechtse content tegengaan. Proactieve interventies worden voornamelijk ingezet door de overheid. Het kan hierbij zowel gaan om generieke interventies (bijvoorbeeld het geven van tips over hoe om te gaan met social media) als om casusgerichte interventies (zoals Stopgesprekken met gebruikers). Deze interventies zijn met name gericht op de eerste fasen van radicalisering, op de ontvankelijkheid voor extremistische content, en op het creëren van een minder extremistische omgeving. Binnen het offline radicaliseringsdomein is veel kennis over de effectiviteit van deze interventies. Reactieve interventies worden voornamelijk ingezet door commerciële instellingen (service providers). Het gaat hier met name om interventies die bepaalde content verwijderen of ontoegankelijk maken, en daarmee worden indirect ook de eerste fasen van radicalisering geadresseerd. Onderzoek naar de effectiviteit van reactieve interventies is nog schaars.

Een uitdaging voor de toekomst is de samenwerking tussen overheden en service providers op het gebied van online extremisme. Beide partners hebben hun eigen grondslag en mogelijkheden om in te grijpen, zowel proactief als reactief, bij extremistische content en de verspreiding daarvan. Samenwerking is dus nodig om op een adequate manier om te gaan met online extremisme, maar die is nog niet optimaal (zie ook (Van Wonderen, et al., 2023). Er zijn echter voorbeelden waarin de outline voor een dergelijke samenwerking zichtbaar wordt. De Christchurch Call is in dit verband een zeer belangrijk moment geweest voor zowel overheden als service providers, deze is genoemd als aanleiding voor een verbeterde samenwerking in alle interviews. De Christchurch Call verwijst naar de gezamenlijke inzet van overheden en technologiebedrijven om online extremisme en terroristische content aan

[26] The War Room | Cobratate
[27] Demonetisering is het onmogelijk maken om geld te verdienen aan bepaalde content. Zie: YouTube demonetization: Protect your videos and earnings in 2023 (uppbeat.io)

te pakken. Het initiatief ontstond als reactie op de aanslagen op moskeeën in Christchurch, Nieuw-Zeeland, op 15 maart 2019, waarbij de dader de aanval live streamde op Facebook. Het doel is om wereldwijd samen te werken aan het elimineren van terroristische en extremistische content online, met respect voor principes van vrije meningsuiting. Het initiatief vertegenwoordigt een wereldwijde inspanning om de verspreiding van extremistische content op internet tegen te gaan en herhaling van de tragische gebeurtenissen van Christchurch te voorkomen. Dit initiatief is het begin geweest van aanzienlijk versterkte (internationale) samenwerking tussen zowel overheden als service providers.

5 Behoeften van de stakeholders rond detectie en interventie

In dit hoofdstuk beschrijven we de door stakeholders gerapporteerde behoeften rond detectie en interventie. Deze behoeften zijn verzameld via interviews en workshopinteractie.

5.1 Behoeften voor detectie

5.1.1 Bewustzijn van nieuwe bedreigingen

Stakeholders geven aan slechts in beperkte mate zich bewust te zijn van nieuwe AI-gebaseerde dreigingsbeelden rond online extremisme. Partijen geven aan dat dit onvolledige beeld belemmerend werkt voor detectie en opvolging.

Voor wat betreft de wetenschappelijke state-of-the-art die relevant is voor het onderwerp radicalisering geeft een meerderheid aan de stand van zaken nauwgezet te volgen. Evenwel blijkt er een discrepantie te bestaan tussen dit kennisniveau en daadwerkelijke detectie en opvolging.

5.1.2 De werkverdeling tussen *recall* en *precision*

Een meerderheid van de stakeholders heeft een voorkeur voor detectiemethoden die een hoge *recall*[28] hebben (met een kans op foutpositieven: onterechte detecties), ten opzichte van methoden met een hoge *precision*, die weer een verhoogde kans hebben op foutnegatieven (ten onrechte gemiste detecties). Enkele stakeholders geven aan dat zij zich niet kunnen veroorloven detecties te missen, en stellen daar een verhoogde menskracht tegenover die detectiealgoritmen controleert op hun uitkomsten.

5.1.3 Uitlegbaarheid

Een grote meerderheid van de stakeholders ziet uitlegbaarheid van detectiealgoritmen als zeer wenselijk, om redenen als politieke integriteit (hierbij verwijzend naar situaties als de toeslagenaffaire), en actiemogelijkheden (interventie en juridische vervolging, waarbij uitlegbaarheid een belangrijke rol kan spelen in de bewijsvoering).

5.1.4 SELP: Societal, Ethical, Legal, Privacy

Een overtuigende meerderheid van de stakeholders voelt zich beknot of gehinderd door vigerende wetgeving rond SELP aspecten. Deze wetgeving staat bijvoorbeeld grootschalige dataverzameling in de weg, en wordt volgens een stakeholder ook verkeerd geïnterpreteerd,

[28] Zie Bijlage A voor een uitleg van de begrippen *recall* en *precision*.

met praktische hindernissen tot gevolg. Ook de traagheid waarmee wetgeving de snelle ontwikkelingen in de AI volgt wordt genoemd als een complicerende factor. Er is behoefte aan uitzonderingsclausules om effectiever te kunnen opereren, ondersteund door nieuwe overlegstructuren tussen stakeholders en overheid.

5.1.5 Samenwerking en experimenteren

Stakeholders geven aan behoefte te hebben aan praktische samenwerking rond detectie en AI – bijvoorbeeld in de vorm van gedeelde lab- of experimenteeromgevingen, met nauwe betrokkenheid van kennispartners met toegepast-wetenschappelijke kennis.

5.2 Behoeften voor interventie

Vanuit de interviews en de workshops zijn vanuit de huidige praktijk en als reactie op de gedeelde kennis uit de literatuur nieuwe behoeften naar voren gekomen die soms zijn ingevuld met nieuwe mogelijke oplossingsrichtingen. Hieronder worden puntsgewijs de belangrijkste behoeften en mogelijke handreikingen zoals die uit de participanten naar voren kwamen gedeeld.

5.2.1 Van fysieke interventies naar aanvullende online interventies

Met name vanuit overheidspartijen wordt de wens geuit om naast het palet van fysieke interventies, waarvan er voorbeelden in hoofdstuk 2 zijn opgesomd, ook meer online interventies te ontwikkelen en hiermee te gaan experimenteren. Denk bijvoorbeeld aan de fysieke STOP-interventie[29] die de politie kan hanteren na detectie van online content die schadelijk of strafbaar is. Hiermee is in een online variant geëxperimenteerd, zoals bijvoorbeeld beschreven in 'interventies jeugdige daders cybercrime" (Oosterwijk, 2017). Hoewel de fysieke STOP-gesprekken onderzocht zijn en bewezen effectief lijken (politie intern rapport), is dat bij online interventies vanuit de overheid nog nauwelijks het geval. Een online interventiecategorie die al langer bestaat is die van online campagnes, door bijvoorbeeld advertenties gericht te plaatsen op specifieke onderwerpen en specifieke doelgroepen. In feite is dit al een goed voorbeeld van publiek-private samenwerking want het vindt plaats in samenwerking met de sociale media platformen. Ten slotte helpt de overheid steeds vaker met online handelingsperspectieven, bijvoorbeeld om de meldingsbereidheid omhoog te krijgen of slachtoffers hulp te bieden. Deze acties variëren van preventie, het beperken van verdere schade tot aan het oplossen van strafbare feiten, waarbij de politie met het openbaar ministerie online opsporingscommunicatie kan inzetten om bijvoorbeeld daders op te sporen.

In de laatste workshop is geconcludeerd dat sociale mediaplatformen en online serviceproviders nu eenmaal meer mogelijkheden hebben om online te interveniëren en overheidspartijen meer kunnen ingrijpen in de fysieke wereld. Het blijft belangrijk om elkaars kracht hierin te onderkennen en benutten. Dat betekent dat als overheidspartijen meer online willen ingrijpen, dit in samenhang met de online providers bezien zal moeten worden, en zij ook proactief kunnen meedenken in overheidsinterventies online. Andersom is het belangrijk dat wereldwijde platforms, wanneer zij meer met lokale partners samenwerken, goed zicht hebben op het landschap/netwerk van partijen die hierin een rol spelen . Verderop in dit hoofdstuk gaan we nader in op de integraliteit van de aanpak.

[29] Cahier 2009-2 Stopreactie-proces_1252b.doc (wodc.nl)

5.2.2 Van reactief naar proactief

Uit de workshopsessies komt naar voren dat overheidspartijen zichzelf en de commerciële partijen als voornamelijk reactief inschatten. In het proactief handelen schiet men nog veel te kort is de gedeelde mening. Dat betekent dat men nu vooral reageert op meldingen of verzoeken van elkaar of van burgers. De partijen geven aan dat capaciteitstekort de belangrijkste oorzaak is: de huidige inzetcapaciteit is nauwelijks genoeg om adequaat reactief te handelen en verhindert additionele inzet op proactief handelen.

De commerciële partijen geven voorbeelden van gewenste interventies die niet alleen gaan over verplichte of gevraagde handelingen, maar ook gericht zijn op vrijwillig en proactief ingrijpen. Hierbij kan gedacht worden aan het geven van voorlichting, het updaten van community richtlijnen, en informatieverstrekking aan partners waar mogelijk of allerlei vrijwillige publiek-private samenwerkingen om het internet veiliger en leefbaar te maken. Hiermee wordt onderscheid gemaakt tussen zogenaamde *upstream* interventies en *downstream* interventies, waarbij 'upstream' vooral gaat over vroegtijdige preventie en 'downstream' over repressie. In sommige literatuur (Schmid, 2020) wordt ook *midstream* nog gebruikt om aan te duidelijk dat er tijdige (*just in time*) in plaats van vroegtijdige interventies plaatsvinden. Het voorkomen van het uploaden van bepaalde inhoud zou hier een voorbeeld van kunnen zijn, zodat het uitvoeren van een extremistische campagne voor voorbereidende handelingen precies op dat moment worden verstoord.

Overheidspartijen hebben aangegeven dat sommige processen in de *downstream* vereenvoudigd zouden kunnen of moeten worden om sneller te handelen. Voorbeelden zijn verzoeken voor contentverwijdering die weliswaar in een uur uitgestuurd kunnen worden, maar soms toch dagen kunnen duren voordat duidelijk is dat het processen van andere overheidspartijen (opsporing of inlichtingen) niet in de weg zit. Zo'n verzoek zou in veel gevallen sneller bij een online dienst of platformbeheerder kunnen liggen is de consensus.

Naast *upstream* of *downstream* als moment van interveniëren maakten de partijen in de discussies ook onderscheid tussen twee routes voor interventies: *bottom-up* en *top-down*. Bottom up interventies komen bijvoorbeeld uit concrete casuïstiek voort, bijvoorbeeld rondom een concreet extremistisch netwerk dat misschien model staat voor andere netwerken of een prototypisch voorbeeld blijkt van een nieuw fenomeen dat dient te worden aangepakt. Top-down interventies komen eerder voort uit strategische analyses waarin nieuwe trends worden waargenomen, of nieuwe technologische mogelijkheden die nieuw gedrag mogelijk zullen maken waarin nieuwe aanpakken nodig zijn. In zowel de bottom-up route als de top-down route is er veel meer proactief ingrijpen mogelijk dan nu, om bredere effecten tegen te gaan. Juist als er informatie over gedeeld kan worden tussen verschillende partners, zodat het er meer sprake is van integraal proactief handelen en ook vroegtijdiger en preventief ingegrepen kan worden om schade te voorkomen in plaats van te repareren.

5.2.3 Gezamenlijke aanpak

Vrijwel alle partijen hebben zowel in de interviews als in de workshops aangegeven meer zicht te willen hebben op wie nu precies wat doet op het gebied van online extremisme. Dit geldt zowel op het gebied van detectie als op het gebied van interventies.

Voor terrorisme is het beeld voor de meeste partijen veel duidelijker. Een gedeelde opvatting is dat terrorisme duidelijker afgebakend is en dat de verantwoordelijkheden daarmee duidelijker belegd zijn.

Niet alle deelnemende partijen hebben een actieve rol op het gebied van detectie of interventie,. Door het in kaart brengen van alle rollen en verantwoordelijkheden die partijen wèl hebben of nemen ontstaat er meer inzicht in het landschap van de aanpak van online extremisme.

Door de gesprekken en workshops is een duidelijker beeld ontstaan van het landschap aan detectie-en interventiemogelijkheden, met een beter begrip van de lacunes in interventies en meer zicht op online extremisme. Bij de Nederlandse overheid is duidelijk dat sommige partijen (zoals opsporingsdiensten) operationeel en casusgericht (vaak persoonsgericht) werken, evenals de ATKM dat heel gericht kan ingrijpen op online situaties (alleen in bij terroristische content en - op termijn - kinderpornografisch materiaal), maar zich niet of heel weinig kan richten op tactische of strategische analyses. Deze taak zou eerder bij veiligheidsdiensten zoals inlichtingen of de NCTV kunnen liggen die samen met andere partners (meer) strategische of tactische trendbeelden kunnen verzamelen en delen. Er wordt tussen veiligheidsdiensten duidelijk onderscheid gemaakt tussen informatie die nodig is voor beleidsvorming en informatie die nodig is voor de operationele uitvoering van veiligheidstaken. Juist hier geven deelnemers aan is er meer samenwerking mogelijk vanuit ieders rol, zoals bijvoorbeeld: vanuit operationele onderzoeken zou vaker een trend in de modus operandi naar boven kunnen komen die (vaker) gedeeld kan worden, terwijl strategische inzichten nieuwe trends identificeren die in Nederland kunnen gaan spelen, en wetenschap die bijdraagt om gefundeerder inzichten te onderzoeken (zoals aard en omvang) waarop beleid kan worden gebaseerd en wellicht onderzoeksjournalistiek die met verrassende nieuwe fenomenen komt die gemist zijn. En zo zijn er meer puzzelstukjes vanuit de diverse rollen die andere partijen vervullen, van repressieve aanpak tot proactief aan preventie werken. Als laatste wordt het politieke belang nog genoemd dat ook van invloed is op hoe de gezamenlijke aanpak verder vormgegeven wordt.

Naast de hiaten in het landschap voor detectie en interventie waar partijen op hebben gewezen als het gaat om toepassing ervan in de praktijk, is er ook een behoefte om processen te versnellen die nu nog (te) veel door bureaucratische processen en risicomijdend gedrag worden belemmerd. Gerapporteerde *best practices* laten zien dat daar waar afdelingen flexibeler capaciteit uitwisselen, met mensen die ervaring hebben opgebouwd bij diverse organisaties, er soepelere samenwerkingen ontstaan waardoor de ruimte die er wel degelijk is voor effectievere samenwerking meer benut lijkt te worden en nieuwe mogelijkheden worden gevonden. Dit is van toepassing op overheidsorganisaties, maar ook op grote commerciële organisaties waarin het werk in divisies of afdelingen is opgedeeld. Bij die laatste zie je bijvoorbeeld dat de afdeling die zich bezig houdt met de monetaire stromen op een platform of de advertenties niet zomaar vanzelf samenwerkt met content moderatieteams of beleidsafdelingen als het om het uitwisselen van kennis over extremisme of meedenken in nieuwe interventies op dit onderwerp gaat. Voor zowel publieke als private partijen blijft het gevoel bestaan dat men nog veel niet weet op het gebied van extremisme, en wat je niet weet kun je ook niet (informatiegestuurd) aanpakken.

Integrale aanpak over alle rechtsdomeinen heen

Alleen door meer afstemming en synchronisatie van interventies kan er een integrale aanpak ontstaan waarin de som der delen effectiever zouden kunnen worden, is hierin de

aanname. Zo is er in Nederland wel een barrièremodel[30] voor politie en openbaar ministerie in de maak voor strafbare online feiten, maar dit is slechts een kleine stap in het grotere geheel aan mogelijke maatregelen om in andere rechtsdomeinen dan alleen strafrecht te interveniëren bij maatschappelijk ongewenste content of gedrag.

Ketenbrede aanpak: publiek en privaat

De ketenbrede aanpak vanuit de overheid is reeds genoemd en die wordt door de overheidspartijen zelf ook herkend. Dat deel van het landschap kan nog beter op elkaar aansluiten en ook de snelheid van samenwerken kan op veel punten beter om effectiever te worden. Het landschap verandert misschien wat, omdat er spelers bijkomen zoals nu de ATKM, maar veel overheidspartijen herijken hun rol om de zoveel jaar om te kijken of de kerntaken nog voldoende effectief kunnen worden uitgevoerd en in samenhang zijn met de omgeving. Voor een ketenbrede aanpak aan de private kant werd door de sociale media partijen benoemd dat de keten veel breder bekeken dient te worden. Het gaat niet alleen om de platformen, maar ook om de hosting providers waar extremistische content te vinden is, om gaming platformen en om veel meer. GIFCT[31] is een voorbeeld waarin deze aanpak al veel breder wordt aangepakt onder private technologiebedrijven, waarin ook een drempel bestaat voor deelname als lid. Een aantal private partijen gebruikt al informatie uit andere omgevingen om zicht te krijgen op cross-platform gebruik. Een voorbeeld is een partij als Discord die kennis uitwisselt met gamingplatform Roblox. Voor grote tech-partijen is inter-platformsamenwerking ook van belang, denk bijvoorbeeld aan een aanbieder als Meta die Instagram, WhatsApp en Facebook als platformen heeft, of Microsoft die Xbox of Linkedin platformen aanbiedt. Wetgeving wordt in alle gevallen als drempel ervaren om meer informatie onderling uit te wisselen en ook de maatschappelijke zorgen hierover. Teveel verantwoordelijkheden bij deze platformen neerleggen zou weer tot maatschappelijke discussies kunnen leiden, want er zijn nog steeds zorgen bij burgers over het wel of niet uitwisselen van gegevens tussen deze platformen. Het beleggen van verantwoordelijkheden om meer informatiegestuurd te werken liggen, zeker als dit cross-platform zou moeten, zeer gevoelig. Daar ligt echter wel een groot deel van de uitdaging. Er zijn voorbeelden van maatregelen om cross-platform gebruik tegen te gaan. Commerciële partijen hebben misschien ook een belang om klanten langer vast te houden op hun platform, maar als dit ook recruteringspatronen, naar bijvoorbeeld kleinere platformen of online plaatsen waarvan bekend is dat er extremisme plaatsvindt, kan tegenhouden is dat misschien ook in maatschappelijk belang.

Naast internetpartijen die internetverkeer op alle lagen voorzien van dienstverlening, en het cross-platformgebruik, zijn er ook nog andere aandachtspunten in een bredere ketenbrede aanpak. Wat met name genoemd wordt is de kleine partijen, waarvan het sterke vermoeden is dat online extremisme daar nauwelijks wordt gedetecteerd of aangepakt. In diverse sectoren heeft dit al aandacht, zoals ‘bad hosters’, maar op het gebied van sociale media platformen verdient dit ook aandacht. Niet dat deze partijen niet welwillend zijn, maar het ontbreekt vaak aan personeel, geld of (technische) middelen om er iets aan te doen, soms omdat het start-ups zijn, soms omdat het partijen zijn in landen waar deze randvoorwaarden niet nodig zijn of niet eenvoudig ingevuld kunnen worden.

Er zijn veel vragen en zorgen over of, hoe en wanneer de DSA (en gerelateerde regelgeving) ook moet gelden voor kleinere partijen. In de tussentijd worden TechAgainstTerrorism[32] en GIFCT door vele partijen als voorbeeld genoemd van hoe ook deze kleinere partijen bediend

[30] Zie voorbeelden van dergelijke modellen op www.barrieremodellen.nl
[31] https://gifct.org/membership/
[32] https://techagainstterrorism.org/home

kunnen worden met kennis en vaardigheidstraining[33], en toolboxes om aan meer detectie en interventie te doen. Grotere technologiepartijen, maar ook de overheidspartijen geven aan graag te willen bijdragen aan meer kennisuitwisseling waar ook deze partijen van kunnen profiteren.

Vertrouwen bouwen

De aanwezige partijen geven aan dat er nog veel nodig is om vertrouwen verder op te bouwen. Onderling in de eigen ketens, maar ook publiek-privaat en zeker internationaal. De deelnemers geven aan dat vertrouwen wordt opgebouwd door positieve ervaringen, en niet door dwang. Informatieuitwisseling over platforms (cross-platform) kan volgens veel deelnemers niet zomaar met overheidsdeelname, omdat eindgebruikers van deze platforms dan mogelijk massaal afhaken.

Het Internet Government Forum (IGF) van de Verenigde Naties wordt in dit verband genoemd als mogelijke partij die hierin een belangrijke rol zou kunnen vervullen. Maar ook concrete initiatieven zoals de GICT of TechAgainstTerrorism bieden succesvolle best practices waarin aan een eco-systeem van partijen wordt gebouwd die op basis van onderling vertrouwen kennis, tools en informatie delen op Europees niveau en wereldwijd. Tussen overheden op nationaal, Europees en wereldwijd niveau is informatiedeling geen eenvoudige opgaaf. Daar moet blijvend aan gewerkt worden door partijen als Europol en Interpol maar ook liaisons om samenwerking en informatie-uitwisseling op verschillende niveau's te organiseren en duurzaam te borgen.

Kennis is een ruilmiddel geworden vanuit de wens naar wederkerigheid, waardoor het ook misbruikt kan worden. Tegelijkertijd is er ook de trend naar meer transparantie, zoals de transparantierapporten over contentmoderatie van de social media platformen als voorbeeld. Kan (meer) transparantie ook als een effectieve interventie beschouwd worden? Stel dat er nog meer transparantie komt over de origine van content, de bronnen bij opsporingsonderzoeken, de verspreidingsnetwerken of de bronnen achter de recrutering? In dit verband is ook gesproken over nieuwe interventies zoals 'shadow banning'[34], waarbij platformen content niet tonen in bepaalde tijdslijnen (afscherming van zichtbaarheid door algoritmes). Dat kan enerzijds een dempend effect hebben op de verspreiding van extremistische content, maar tegelijkertijd is de vraag hoe transparant deze praktijk voor zowel eindgebruikers als overheid is. En wat zijn de neveneffecten: wordt gedrag sneller extremistischer misschien? Wie bepaalt er wie wat ziet online? Er is onvoldoende zicht op (neven)effecten van allerlei maatregelen in het geheel en als het onbekend is hoe maatregelen werken kan het ook moeilijker worden om effectief samen te werken.

Privaat-privaat, publiek-publiek en publiek-privaat

De belangen van zowel overheden als commerciële partijen spelen in deze gezamenlijke aanpak wel een belangrijke rol. Om deze reden moet goed vastgesteld worden welke kennis of informatie gedeeld kan en moet worden voor een optimale gezamenlijke afstemming. Zowel in de commerciële sector als aan de overheidskant worden op dat vlak meer kansen gezien. Voor sommige samenwerkingen is het simpelweg niet handig of mogelijk om deze samenwerking publiek-privaat te doen, zeker ook gezien grote verschillen in de diverse landen. Het onderling vertrouwen en gelijkwaardigheid tussen publieke en private partners is hierin niet de enige uitdaging – ook legitimiteit, mededinging en vele andere dilemma's spelen een rol die goed afgewogen moet worden. Als voorbeeld is gesproken over het publiek-privaat, publiek-publiek of privaat-privaat delen van symbolen of (hashed) video's.

[33] https://gifct.org/year-three-working-groups/
[34] Zie https://en.wikipedia.org/wiki/Shadow_banning

Voor elk van deze toepassingen is wat te zeggen, maar ze hebben een andere uitwerking. Daarnaast is de vraag op welk niveau informatie wordt uitgewisseld. Een voorbeeld is het domein van Cyber Intelligence, waarin op operationeel niveau bijvoorbeeld IP-adressen uitgewisseld kunnen worden van potentiële dadergroepen om direct in te kunnen ingrijpen. Op tactisch niveau wordt er ook informatie uitgewisseld over bijvoorbeeld een nieuwe vorm van *ransomware* die is gesignaleerd of zou kunnen ontstaan door mutatie van een virus. Op tactisch niveau wordt dan kennis en kunde verbeterd en kunnen nieuwe maatregelen getroffen worden voor betere detectie of interventie. Op strategisch niveau zou informatie uitgewisseld kunnen worden over welke aanvalsvormen er zijn geweest en hoe effectief de integrale aanpak ervan was, door bijvoorbeeld KPI's op te stellen en hiermee te sturen op detectie of interventie en deze informatie op hoofdlijnen met elkaar te delen, zonder dat er privacygevoelige informatie gedeeld hoeft te worden.

Er is veel meer mogelijk op het gebied van samenwerking in de sectoren en tussen overheidspartijen. Maar er worden daarnaast ook meer kansen voor publiek-private samenwerking gezien. De rol van onafhankelijke partijen die kennis en informatie delen kan hierin als katalysator werken. Wat overheden of bedrijven niet zomaar kunnen doen of aan informatie kunnen delen is soms wel mogelijk voor onafhankelijke non-of not-for-profit organisaties die vanuit andere grondslagen en belangen werken, zoals bijvoorbeeld een NGO als TechAgainstTerrorism, GIFCT, het Global Network on Extremism & Technology35 ,of een onafhankelijk kennisinstituut als TNO. Ook onafhankelijke ondersteuning van samenwerking en kennisdeling werd door de bevraagde stakeholders expliciet genoemd: de interviews die TNO afnam en de workshops die werden georganiseerd in het kader van dit onderzoek werden door participanten als typisch voorbeeld genoemd.

Naast lokale en Europese ook verdere internationale afstemming
De coördinatie vanuit de EU in afstemming met de VS en andere belangrijke landen (zoals bijvoorbeeld Australië waar contacten mee zijn opgebouwd) op het gebied van online extremisme wordt ook genoemd als verbeterpunt.

5.2.4 Uitbreiding met andere typen partners

Vanuit de betrokken stakeholders wordt een behoefte geuit om meer samen te werken met sociale partners (zoals in de gezondheidszorg of onderwijs) en ook sociale wetenschap over wat er aan trends aan zou kunnen komen. Vragen als "Hoe leven jongeren nu online?" zeggen misschien wat over toekomstige vormen van online gebruik en misbruik van online media in relatie tot extremisme. Overheidsinstanties geven al aan op wijkniveau samen te werken met sociale wijkteams en op rijksniveau wordt al samengewerkt met wetenschap. Participanten geven aan dat intensivering en uitbreiding van deze vormen van samenwerking gewenst is. Ook enkele platformen geven aan lokaler samen te werken of te willen werken met dergelijke instanties, variërend van NGO's (zoals bijvoorbeeld TechAgainstTerrorism) tot wetenschap. Met name de platformen en partijen die wereldwijd opereren geven aan dat er veel verschillen zijn tussen de online omgevingen, maar ook tussen de landen en doelgroepen. Meer focus, zoals de focus op specifieke groepen jongeren die sommige partijen hanteren, helpt bij het gerichter kennis uitwisselen en het afstemmen van interventies.

Daarnaast is er een andere belangrijke reden om (meer) met andere typen partners uit het sociale domein te gaan samenwerken. Dit sluit aan bij de eerder genoemde wens om de aandacht ook meer te verschuiven van reactieve interventies naar proactieve interventies en

[35] https://gnet-research.org/resources/insights/

preventie. De huidige status quo valt logisch te verklaren uit het feit dat men op nationaal en Europees niveau beleid gestart is met de nieuwe wettelijke kaders en regelgeving, en vanuit de nationale veiligheidspartijen van de overheid rollen pakt vanuit de kerntaak in veiligheid (met elk een grondslag en mandaten) die klassiek repressief is. Vanuit deze stappen die nu een aantal jaren gemaakt zijn, lijkt het een logische wens om als vervolgstap meer aandacht te vragen voor preventie en meer aan de voorkant van het extremismeprobleem online te komen. En juist hier spelen publieke en private partijen uit het sociale domein een belangrijke rol. Sommige deelnemers vergeleken het met een voetbalteam, waarbij als de achterhoede van veiligheidspartijen met een meer repressieve aanpak en informatiepositie goed kunnen samenwerken met elkaar en de private internetsector, dan moet de voorhoede gevormd worden door sociale partners. Bij veel van deze sociale partners is ook een beweging te zien naar meer publiek-private samenwerking omdat sociale hulpverleners in wijken de jeugd bijvoorbeeld niet langer alleen in de fysieke wereld helpen.

6 Conclusies en aanbevelingen

6.1 Inleiding

In de conclusie geven we aan de hand van de informatie verzameld uit de interviews, het literatuuronderzoek en de workshops een aantal strategische aanbevelingen. Deze aanbevelingen omvatten de noodzaak voor een verbeterde samenwerking en een expertise die zowel dieper ingaat op het onderwerp van de psychologie en sociologie van radicalisering als de technieken die online gebruikt worden.

6.2 Nieuwe uitdagingen en kansen voor detectie

Generatieve AI De opkomst van generatieve AI biedt zowel nieuwe kansen als bedreigingen. Kansen zijn er op het gebied van de inzetbaarheid van bijvoorbeeld Large Language Modellen als detectoren van radicale of extremistische content. Een mogelijk operationeel voordeel van deze modellen in een praktische workflow bestaat eruit dat deze modellen, na toepasselijke prompting of finetuning, in staat zijn NLP-analyses uit te voeren (zoals de detectie van extremisme in tekst), en daarbij een mens-leesbare uitleg kunnen produceren. Ook kunnen deze modellen mogelijk worden ingezet voor automatische contentmoderatie, zoals het herformuleren van radicale of extremistische teksten, of het formuleren van *counter narratives*.

Er vindt momenteel veel onderzoek plaats naar de prestaties van grote taalmodellen als NLP-analysetools. Systematisch onderzoek naar gebruik en kwaliteit van deze modellen in de context van online extremisme en radicalisering ontbreekt vooralsnog. Daarnaast dient te worden onderzocht hoe -in algemenere zin- Generatieve AI op een communicatieve, interactieve manier kan worden ingezet in detectie-workflows, hoe deze modellen lokaal en veilig kunnen worden gebruikt, en hoe ze kunnen worden aangesloten op bestaande kennisbronnen, tools en databases. Daarbij dient de rol van de menselijke analist goed onderzocht te worden. Technologie als LLMs biedt een geheel nieuwe vorm van communicatieve interactie met detectiegerichte tools, met voordelen (zoals betere uitleg en interpreteerbaee onderbouwing) als nadelen (zoals bias en de kans op onjuiste informatie). Deze voor- en nadelen onderstrepen het belang van goed menselijk toezicht, en het scheppen van de gunstige en juridisch/ethisch acceptabele of zelfs verplichte randvoorwaarden voor de inzet van dit soort technologie.
De tech-industrie biedt in toenemende mate veilige, lokale omgevingen voor Generatieve AI (zoals lokale Azure-cloudoplossingen van Microsoft waarbinnen GPT-modellen bijgetraind en gebruikt kunnen worden), en dat er steeds meer mogelijkheden ontstaan om met name grote taalmodellen te laten communiceren met externe tools en databases[36]. Verder ontstaan met ontwikkelingen als de nieuwe GPT-store van OpenAI[37] mogelijkheden om

[36] Via zogeheten *API calls*.
[37] https://openai.com/blog/introducing-gpts

kleine, gespecialiseerde AI-assistenten te trainen en veilig binnen een organisatie te gebruiken. Het op een verantwoorde, veilige manier inzetten van dit soort technologie in de veiligheidssector, in dit geval voor detectie en geautomatiseerde contentmoderatie, komt daarmee een stap dichterbij. De potentie om met Generatieve AI op grote schaal radicale of extremistische content te produceren, geautomatiseerd en op maat voor bepaalde groeperingen (qua taal, stijl of onderwerp) vormt een nieuwe bedreiging voor de verspreiding en detectie van online extremistisch gedachtegoed. Hiervan zijn diverse voorbeelden in de recente literatuur en de praktijk aangetroffen, zoals de aanwezigheid van deep fake beelden van het Israël-Hamas conflict van 2023 in de commerciële beeldcatalogus van Adobe[38]. De detectie van dit soort zeer natuurlijk ogend materiaal vraagt om nieuwe detectietechnieken.

Hyperrealism Recent onderzoek (Miller *et al.* , 2023) wijst op het gevaar van AI Hyperrealism– het verschijnsel dat mensen door AI gegenereerde content natuurlijker vinden dan natuurlijke content (in het onderzoek van Miller *et al.* gaat het om het onderscheid tussen natuurlijke en synthetische gezichten). Dit onderzoek laat zien dat mensen weliswaar in staat zijn de beoordelingscriteria voor het onderscheid tussen natuurlijke en synthetische content kunnen benoemen, maar dat ze die criteria verkeerd toepassen, Verrassend genoeg bleken machine learning-modellen voor gezichtsbeoordeling die gebruik maken van deze criteria (en ze wèl goed toepasten) nauwkeuriger dan de mens in het onderscheiden van natuurlijke versus synthetische content. Dit biedt hoop voor een nieuwe, mens-geïnformeerde generatie van op beeld gerichte detectietools. Dergelijke experimenten kunnen ook op door LLM en mensen geproduceerde teksten worden toegepast. Mogelijk onderscheiden mensen subtiele aspecten aan synthetische tekst die behulpzaam zijn voor het geautomatiseerd onderscheid maken tussen natuurlijke en door AI gegenereerde tekst.

Adversarial attacks

Naarmate de detectie van online radicalisering geavanceerder wordt, wordt het voor geradicaliseerde entiteiten interessanter om eventueel publiek benaderbare detectiemodellen die door overheidsinstanties worden gebruikt te ondermijnen met dit soort *adversarial attacks* (zie Bijlage A). Detectie van *black box attacks* op publieke AI services is van belang als overheden en andere stakeholders besluiten dit soort services in te zetten als detectors. Li *et al.* (2022) presenteren een techniek om query-gebaseerde blackbox attacks vroegtijdig te herkennen.

Efemerische content

Tenslotte zijn er steeds meer platformen, zoals pastebin, Discord of Snapchat, waarbij in real-time informatie gedeeld wordt die slechts tijdelijk beschikbaar is (vluchtige, of *efemerische* content). Hierbij is het van belang dat overheidsinstanties live informatie kunnen aftappen waar modellen direct op toegepast kunnen worden en niet afhankelijk zijn van het achteraf verzamelen van datasets. Hierover is echter wel veel maatschappelijke discussie[39], getuige een van de oplossingsrichtingen die in een nieuw wetsvoorstel van de Europese Unie wordt voorgesteld om zogenaamde client-side scanning te kunnen uitvoeren. Dit betreft het controleren van berichten op de smartphone zelf door aanbieders van online diensten, waarbij afbeeldingen, video's en teksten worden vergeleken met een database en, als er een match is, de politie kan worden ingeschakeld.

[38] https://www.businessinsider.com/adobe-stock-ai-generated-images-israel-hamas-war-2023-11 (December 2023).
[39] cf5290564050b55b72b93c83231be68a.pdf (euobserver.com)

6.3 Nieuwe uitdagingen en kansen voor interventie

Publiek-private samenwerking rond afstemming van online en offline interventies is nog pril. De overheid wil meer online interventies verkennen in samenwerkingsverband, terwijl online platformen en service providers zoeken naar wat er lokaal en in de fysieke contacten meer gedaan kan worden met specifieke doelgroepen of problematiek. Daarnaast is er ook vanuit de platformen een beweging om meer aan preventie van schadelijke content en preventie van verspreiding te doen in plaats van alleen contentmoderatie, waarop de bedrijfsprocessen nu door wet- en regelgeving op veel plaatsen al zijn ingeregeld.
Een meer integrale aanpak waarbij ook steeds meer sociale partners die zich op preventie richten een rol spelen is wenselijk, naast de meer reactieve aanpak vanuit veiligheidsorganisaties. Meer afstemming (of synchronisatie) in maatregelen om tot een meer integrale aanpak te komen wordt als zeer wenselijk gezien.

Er zijn recentelijk ook nieuwe organisaties ontstaan die een specifiekere rol opnemen in het publiek-private speelveld. De ATKM in Nederland en GIFCT en TechAgainstTerrorism internationaal zijn daar mooie voorbeelden van. Interventies coördineren, onderlinge processen verbeteren, kennisdeling, het op weg helpen van (nieuwe) platformen en (samen) ontwikkelen van technologie zijn voorbeelden van taken die deze partijen op zich nemen.

Nieuwe vormen van samenwerking ontstaan ook vanuit private en publieke partijen zelf, waarbij in private-private publieke-publieke en publieke-private interacties hun eigen samenwerkingsmogelijkheden verkennen. Wet- en regelgeving maar ook imago of legitimiteit en onderling vertrouwen spelen een belangrijke rol bij de haalbaarheid van dit soort samenwerking.

Op diverse niveaus, van operationeel tot tactisch en strategisch kan er meer uitgewisseld worden aan ervaringen en nieuwe aanpakken, want onderzoek naar de effectiviteit van interventies is nog schaars.

6.4 Samenwerking blijven stimuleren

Er moet meer focus komen op samenwerking. De NCTV zou hierin een faciliterende rol kunnen spelen door een team van experts vanuit verschillende diensten samen te stellen en te ondersteunen. Diensten die in ieder geval betrokken zouden moeten zijn, zijn een innovatieteam of een landelijk georganiseerde eenheid.

Daarnaast zijn er ook samenwerkingen in het publieke domein gewenst. Door meer te focussen op online gedrag, bijvoorbeeld met behulp van burgerschapslessen, kan proactief gewerkt worden aan het voorkomen en verminderen van verspreiding van extremistische content.

6.5 Vervolgonderzoek

Deze studie heeft vanuit de literatuur en de praktijk trends en ontwikkelingen in kaart gebracht in de fenomenologie van online radicalisering en extremisme, de detectie van online radicale en extremistische content, en interventies die de verspreiding van deze content moeten tegengaan. Deze onderwerpen zijn in relatieve isolatie bestudeerd. Vervolgonderzoek dient zich te richten op de vraag wat de praktische operationalisering van

deze onderwerpen is, en welke kruisverbanden er tussen deze onderwerpen bestaan of kunnen worden gelegd. Denkbare onderzoeksvragen hierbij zijn:

1. Hoe verhouden zich huidige state-of-the-art detectiemethoden tot de nieuwe AI-gebaseerde dreigingsbeelden voor het produceren van radicale of extremistische content? Hoe herkennen we AI-gegenereerde content?
2. Wat zijn geschikte metrieken om de effectiviteit van online interventies betrouwbaar in te schatten (kwalitatief en kwantitatief)?
3. Hoe betrekken we online gedragsanalyse en psychometrie bij content-gebaseerde detectie?

Referenties

Ahmad, S., Zubair Asghar, M., Alotaibi, F. & Awan, I. (2019). Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. Human-centric Computing and Information Sciences, 9, 1-23.

Alexander, J. (2018, Maart 10). *The Yellow $: a comprehensive history of demonetization and YouTube's war with Creators.* Opgehaald van Polygon: https://www.polygon.com/2018/5/10/17268102/youtube-demonetization-pewdiepie-logan-paul-casey-neistat-philip-defranco

Bates, J. (2017). The Politics of Data Friction. *Journal of Documentation*, pp. 412-429. doi:10.1108/JD-05-2017-0080

Benigni, M., Joseph, K. & Carley, K. (2017). Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter. *PloS one, 12(12)*.

Bermingham, A., Conway, M., McInerney, L., O'Hare, N. & Smeaton, A. (2009). Combining Social Network Analysis and Sentiment Analysis to Explore the Potential for Online Radicalisation. *2009 International Conference on Advances in Social Network Analysis and Mining*, 231-236.

Bhui, K., Dinos, S. & Jones, E. (2012). Psychological process and pathways to radicalization. . *Journal of Bioterrorism & Biodefense, 5*, 1-5.

Bolet, D. & Foos, F. (2023). Media platforming and the normalization of extreme right views. *URPP Equality of Opportunity Discussion Paper series no. 22*.

Bond, M. (2014, September 10). *Why westerners are driven to join the jihadist fight.* Opgehaald van New Scientist: https://www.newscientist.com/article/mg22329861-700-why-westerners-are-driven-to-join-the-jihadist-fight/#.VQLIAvysX_E

Borum, R. (2012). Radicalization into violent extremism II: A review of conceptual models and empirical research. *Journal of Strategic Security, 4(4)*, 37–62.

Braddock, K. (2022, November 25). Vaccinating Against Hate: Using Attidudinal Inoculation to Confer Resistance to Persuasion by Extremist Propaganda. *Terrorism and Political Violence*, pp. 240-262. doi:10.1080/09546553.2019.1693370

Buerger, C. (2022, April 27). Counterspeech: A Literature Review. *SSRN*. doi:10.2139/ssrn.4066882

Cialdini, R. B. (2001). The science of persuasion. *Scientific American, 284(2)*, 76-81.

Cunningham, K. (2007). Countering Female Terrorism. *Studies in Conflict and Terrorism 30(2)*, 113-129.

Desmarais, S.L., Simons-Rudolph, J., Brugh, C.S., Schilling, E. & Hoggan, C. (2017). The state of scientific knowledge regarding factors associated with terrorism. *Journal of Threat Assessment and Management, 4(4), 180.*., 180.

Dodd, V. (2023, January 26). *Large rise in men referred to Prevent over women-hating incel ideology.* Opgehaald van The Guardian: https://www.theguardian.com/uk-news/2023/jan/26/large-rise-in-men-referred-to-prevent-over-women-hating-incel-ideology

Ecker, U.K., Butler, L.H. & Hamby, A. (2020). You don't have to tell a story! A registered report testing the effectiveness of narrative versus non-narrative misinformation corrections . *Cognitive Research: Principles and Implications, 5(1)*.

Ecker, U.K., Sanderson, J.A., McIlhiney, P., Rowsell, J.J., Quekett, H.L., Brown, G.D. & Lewandowsky, S. (2022). Combining refutations and social norms increases belief change. *Quarterly Journal of Experimental Psychology*.

Elshout, T. v. (2023, Augustus 31). Tientallen Vrouwen uitgebuit Via Online Platform Andrew Tate. *RTL Nieuws*. Opgehaald van https://www.rtlnieuws.nl/nieuws/buitenland/artikel/5404984/andrew-tate-war-room-vrouwen-uitbuiting-verkrachting-online

Epstein, Z., Berinsky, A.J., Cole, R., Gully, A., Pennycook, G. & Rand, D. G. (2021). Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. . *Harvard Kennedy School Misinformation Review, 2(3).*

Feddes, A.R., Nickolson, L. & Doosje, B. (2015). *Trigger variabelen in het radicaliseringsproces.* Expertise-unit Sociale Stabiliteit/Universiteit van Amsterdam.

Floridi, L. (2015). The Online Manifesto: Being Human in a Hyperconnected Era. Cham. doi:10.1007/978-3-319-04093-6

Garland, J., Ghazi-Zahedi, K., Young, J.G., Hébert-Dufresne, L. & Galesic, M. (2022). Impact and dynamics of hate and counter speech online. *EPJ Data Science*. doi:10.1140/epjds/s13688-021-00314-6

Ghosh, S. & Stocking, G. (2023, May 19). On Alternative Social Media Sites, Many Prominent Accounts seek Financial Support from Audiences. Opgehaald van https://www.pewresearch.org/short-reads/2023/05/19/on-alternative-social-media-sites-many-prominent-accounts-seek-financial-support-from-audiences/?utm_source=Pew+Research+Center&utm_campaign=1a65ddacb9-Internet-Science_2023_05_25&utm_medium=email&utm_ter

Gialampoukidis, I., Kalpakis, G., Tsikrika, T., Vrochidis, S. & Kompatsiaris, I. (2016). Key player identification in terrorism-related social media networks using centrality measures. *European Intelligence and Security Informatics Conference (EISIC)*, 112-115.

Herath, C. & Whittaker, J. (2023). Online radicalisation: Moving beyond a simple dichotomy. *Terrorism and political violence, 35(5)*, 1027-1048.

Jamil, M., Luqman, S., Pais, S., Cordeiro, J. & Dias, G. (2022). Detection of extreme sentiments on social networks with BERT. *Social Network Analysis and Mining, 12(1)*.

Jigsaw. (2023, July 14). *6 Insights Into How Hate & Violent Extremism are Evolving Online.* Opgehaald van Medium: https://medium.com/jigsaw/6-insights-into-how-hate-violent-extremism-are-evolving-online-f8faf4521398

Julia Ebner, C.K. (2023, July 25). Assessing Violence Risk among Far-Right Extremists: A New Role for Natural Language Processing. *Terrorism and Political Violence*. doi:https://doi.org/10.1080/09546553.2023.2236222

Kahneman, D. (2012). Two systems in the mind. Bulletin of the American Academy of Arts and Sciences, 65(2), 55-59.

Kaur, A., Kaur Saini, J. & Bansal, D. (2019). Detecting Radical Text over Online Media using Deep Learning. *arXiv preprint arXiv:1907.12368*.

Klausen, J., Marks, C. & Zaman, T. (2018). Finding extremists in online social networks. *Operations Research, 66(4)*, 957-976.

Kozyreva, A., Lorenz-Spreen, P., Herzog, S., Ecker, U., Lewandowsky, S., Hertwig, R. & et al. (2023). Toolbox of interventions against online misinformation – Version 2.0. Online supplement for the expert review project "Toolbox of Interventions Against. Opgehaald van https://interventionstoolbox.mpib-berlin.mpg.de/table_concept.html

Kruglanski, A.W. (2018). Violent radicalism and the psychology of prepossession. *Social Psychological Bulletin 13(4)*, e27449.

Kruglanski, A.W., Gelfand, M.J., Bélanger, J.J., Sheveland, A., Hetiarachchi, M. & Gunaratna, R. (2014). The psychology of radicalization and deradicalization: How significance quest impacts violent extremism. *Political Psychology, 35*, 69–93.

Leidig, E. (2021). *We Are Worth Fighting For: Women in Far-Right Extremism.* ICCT. Opgehaald van https://www.icct.nl/publication/we-are-worth-fighting-women-far-right-extremism

Leidig, E. (2023). *The Women of the Far Right.* New York: Columbia University Press.

López-Sánchez, D., Corchado, J. & González Arrieta, A. (2018). Dynamic Detection of Radical Profiles in Social Networks Using Image Feature Descriptors and a Case-Based Reasoning Methodology. *Case-Based Reasoning Research and Development: 26th International Conference, ICCBR 2018, Stockholm, Sweden, July 9-12, 2018, Proceedings 26*, 219-232.

McCann, E.M.-I. (2023, March). The Link between Misinformation and Radicalisation: Current Knowledge and Areas for Future Inquiry. *Perspectives on Terrorism*, pp. 33-49. Opgehaald van https://www.jstor.org/stable/27209215

Mitts, T. (2021, May 6). Countering Violent Extremism and Radical Rhetoric. *International Organization*, pp. 251-272.

Moghaddam, F.M. (2005). The staircase to terrorism: A psychological exploration. *American Psychologist, 60(2)*, 161–169.

Moussaoui, M., Zaghdoud, M. & Akaichi, J. (2019). A possibilistic framework for the detection of terrorism-related Twitter communities in social media. *Concurrency and Computation: Practice and Experience 31.13*.

Murray, C. (2023, Augustus 31). What We Know About Andrew Tate's 'War Room' - As Report Alleges Global Network to Exploit Women. *Forbes*. Opgehaald van https://www.forbes.com/sites/conormurray/2023/08/31/what-we-know-about-andrew-tates-war-room-as-report-alleges-global-network-to-exploit-women/

Nickolson, L., Bergen, N. V., Feddes, A., Mann, L. & Doosje, B. (2021). *Extremistisch denken en doen.* WODC.

Nouh, M., Nurse, J. & Goldsmith, M. (2019). Understanding the radical mind: Identifying signals to detect extremist content on Twitter. *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 98-103.

Pennebaker, J., Boyd, R., Jordan, K. & Blackburn, K. (2015). The development and psychometric properties of LIWC2015.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G. & Rand, D.G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science, 31(7)*, 770–780.

Piazza, J.A. (2006). Rooted in poverty? Terrorism, poor economic development, and social cleavages. *Terrorism & Political Violence, 18(1)*, 159–177.

Rajendran, A., Sahithi, V., Gupta, C., Yadav, M., Ahirrao, S., Kotecha, K., . . . Alhammad, S. (2022). Detecting Extremism on Twitter During U.S. Capitol Riot Using Deep Learning Techniques. *IEEE Access, 10*, 133052-133077.

Robin O'Luanaigh, H. R. (2023). *The Right Fit: How Active Club Propaganda Attracts Women to the Far-Right.* Global Network on Extremism and Technology. Opgehaald van https://gnet-research.org/2023/12/05/the-right-fit-how-active-club-propaganda-attracts-women-to-the-far-right/

Roozenbeek, J., Van der Linden, S., Goldberg, B., Rathje, S. & Lewandowsky, S. (2022, August 24). Psychological inoculation improves resilience against misinformation on social media. *Science Advances*. doi:10.1126/sciadv.abo6254

Sageman, M. (2004). *Understanding terror networks.* University of Pennsylvania Press.

Saleh, N.F., Roozenbeek, J., Makki, F.A., McClanahan, W.P. & Van der Linden, S. (2021, February 01). Active Inoculation Boosts Attitudinal Resistance against Extremist Persuasion Techniques: a Novel Approach towards the Prevention of Violent Extremism. *Behavioural Public Policy*, pp. 1-24. doi:10.1017/bpp.2020.60

Sarah L. Carthy, C.B. (2023, August 12). Counter-Narratives for the Prevention of Violent Radicalisation: A Systematic Review of Targeted Interventions. *Campbell Systematic Reviews*. doi:https://doi.org/10.1002/cl2.1106

Schmid, A.P. (2013). Radicalization, de-radicalization, counter-radicalization: A conceptual discussion and literature review. *ICCT Research Paper, 97(1)*, 22.

Schmid, P. & Betsch, C. (2019). Effective strategies for rebutting science denialism in public discussions. *Nature Human Behaviour, 3(9)*, 931-939.

Silke, A. (2008). Holy warriors: Exploring the psychological processes of jihadi radicalization. *European Journal of Criminology, 5(1)*, 99–123.

Sterkenburg, N. (2021). Maar dat mag je niet zeggen: De nieuwe generatie radicaal- en extreemrechts in Nederland. Das Mag.

Theodosiadou, O., Pantelidou, K., Bastas, N., Chatzakou, D., Tsikrika, T., Vrochidis, S. & Kompatsiaris, I. (2021). Change Point Detection in Terrorism-Related Online Content Using Deep Learning Derived Indicators. *Information, 12(7)*.

Thomas, D.R. & Wahedi, L.A. (2023, June 5). *Disrupting hate: The effect of deplatforming hate organizations on their online audience.* Opgehaald van PNAS: https://www.pnas.org/doi/10.1073/pnas.2214080120#supplementary-materials

Van den Berg, H. & Van Hemert, D. A. (2021). Becoming a violent extremist: a General Need and Affect model of psychological variables. *Behavioral Sciences of Terrorism and Political Aggression, 15(4)*, 429-447.

Van Wonderen, R. & Peeters-Osseyran, M. (2022). *Werken aan weerbaarheid tegen desinformatie en een eenzijdige meningsvorming.* Utrecht: Verweij-Jonker Instituut/RadarAdvies.

Van Wonderen, R., Burggraaff, D., Ganpat, S., Beek, V., Cauberghs, O. & Keijzer, M. (2023). *Rechtsextremisme op sociale mediaplatforms?* Utrecht: Verwey-Jonker Instituut. Opgehaald van https://repository.wodc.nl/handle/20.500.12832/3304

Vergani, M., Iqbal, M., Ilbahar, E. & Barton, G. (2018). The three Ps of radicalization: Push, pull and personal. A systematic scoping review of the scientific evidence about radicalization into violent extremism. *Studies in Conflict & Terrorism, 43(10)*, 1-32.

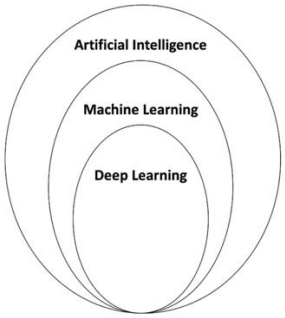
Wells, G., Romhanji, A., Reitman, J.G., Gardner, R., Squire, K. & Steinkuehler, C. (2023). Right-Wing Extremism in Mainstream Games: A Review of the Literature. *Games and Culture, 0*, 1-24.

Whittaker, J. (2022). Rethinking Online Radicalization. *Perspectives on Terrorism, 16(4)*, 27-40.

Whittaker, J. (2023). *Online Radicalisation: What We Know.* European Commission. doi:10.13140/RG.2.2.33102.64322

Bijlage A
Machine learning

Machine Learning-technieken worden veelvuldig gebruikt in het domein van NLP (Natural Language Processing) om informatie te extraheren uit tekst. In de basis is het idee dat een mens handmatig zogenaamde *features* selecteert en die meegeeft aan een model. Het model leert dan uit voorbeelden een complexe functie en past die toe om voorspellingen te doen op niet eerder geziene data.[40] Zoals in Figuur A.1 te zien is, maakt Machine Learning onderdeel uit van Artificiële intelligentie, en is Deep Learning weer een specifiek subtype van Machine Learning.



Figuur A.1: Diagram dat de verhoudingen tussen AI, Machine Learning en Deep Learning weergeeft.

Data-representatie Machine learning-modellen maken gebruik van specifieke representaties van hun training- en input data, en leren het verband tussen die representaties en uitkomstvariabelen (labels). Training- en testdata worden doorgaans door menselijke annotators voorbeweekt en gelabeld. Een voorbeeld is het domein van *sentiment-analyse*. Hierbij leert een model te beoordelen of een tekst, zoals een recensie over een film, een positief of negatief oordeel (of sentiment) bevat. Naast het specificeren van de woordelijke inhoud van zulke teksten, en het (automatisch) statistisch wegen van woorden, kunnen ook andere, minder voor de hand liggende aspecten opgenomen in de tekstrepresentaties, zoals het gebruik van uitroeptekens of de lengte van de tekst. Vaak worden ook zogenaamde n-*grams* gebruikt als features: combinaties van een vast aantal (*n*) woorden uit de tekst, afgeleid door een venster met *n* cellen over een tekst te schuiven,

Een voorbeeld, met *n=3*, is:

- Dit onderzoek richt zich op online extremisme
  [Dit, onderzoek, richt], [onderzoek, richt, zich], [richt, zich, op], [zich, op,online], [op, online, extremisme].

[40] Modellen die vaak gebruikt worden in het NLP-domein zijn onder andere *Logistic Regression*, *Naïve Bayes*, *Random Forests*, *Support Vector Machines* en K-*Nearest Neighbours*. De details van hoe deze modellen werken zijn buiten dit rapport gelaten.

Deze woordgroepen kunnen informatiever zijn dan losse woorden (het woord 'fantastisch' in de woordgroep 'is echt fantastisch' duidt iets heel anders aan dan het woord fantastisch in de woordgroep 'zeker niet fantastisch').

Andere, veelgebruikte features bestaan uit *character* n*-grams*, waarbij niet groepen woorden maar opeenvolgende groepen letters (karakters) worden gebruikt. Een voorbeeld is

- extremisme => {ext, xtr, tre, rem, emi, mis ,ism, sme}

Deze representaties blijken allerlei subtiele syntactische informatie te bevatten, zeker als ze over woordgrenzen heen worden toegepast, en zich uitstrekken over opeenvolgende woorden. Ook reduceren ze problematiek rond spellingsvarianten en tekstuele ruis, doordat ze woorden opknippen in kleine stukjes. Het aanbieden van gelabelde teksten die zijn voorbewerkt met deze feature-representaties stelt modellen in staat op basis van deze "micro-syntactische" features te voorspellen welke recensies positief zijn en welke negatief. Daarbij is het de gewoonte verschillende feature-representaties te combineren.

Deep learning Deep learning is een vorm van machine learning gebaseerd op neurale netwerken met een complexe ("diepe") interne structuur. Deze structuur kan bijvoorbeeld bestaan uit een grote hoeveelheid met elkaar verbonden lagen die allemaal samenwerken om input te analyseren, of uit een complex samenspel van temporele en spatiële analyses (bijvoorbeeld bij het analyseren van tijdsreeksen, teksten of videos). Veel deep learning-modellen kennen tijdens hun trainingsfase zelf *vectorrepresentaties* toe aan hun data, en optimaliseren deze representaties gaandeweg gedurende hun training. Vectoren zijn lijsten met nummers en worden zo berekend dat woorden die een soortgelijke betekenis hebben, ook veel overeenkomsten vertonen in hun numerieke representatie. Voor tekstuele data beschrijven deze representaties allerlei contextuele informatie van woorden[41]. Om deze reden wordt *deep learning* ook wel *representation learning* genoemd. Met name Large Language Models berekenen hooginformatieve contextuele vectoren.

Deep learning-modellen kunnen tijdens het aanleren van hun vector-representaties gekoppeld worden aan een extra leertaak, een zogeheten *downstream taak*. Dit zal tot specifieke vectorrepresentaties leiden die geoptimaliseerd zijn voor de downstream taak. Als zo'n downstream taak bijvoorbeeld bestaat uit sentiment-analyse, en een Deep learning-model leert dat de woorden 'fantastisch' en 'geweldig' goede voorspellers zijn voor positief sentiment, dan zullen de vectoren van deze woorden op elkaar gaan lijken. Hierbij moet worden opgemerkt dat een deel van de gelijkenis tussen deze vectorrepresentaties gebaseerd is op het delen van gelijke woordcontext tussen woorden. Dit betekent dat zonder de fine-tuning van de sentiment downstream taak een vector als "waardeloos" ook sterk op de vector van "fantastisch" zal lijken – deze woorden komen nu eenmaal vaak in dezelfde context voor ("een waardeloze film", "een fantastische film").

Een LSTM is een neuraal netwerk dat, zoals de naam suggereert, een geheugen heeft dat is opgedeeld in een korte- en een lange termijngeheugen . Een LSTM leert zelf bepaalde informatie te onthouden of benadrukken en andere informatie weg te drukken of te vergeten. Voor toepassingen in het NLP-domein geldt dat het netwerk tijdens de computatie langere reeksen woorden 'onthoudt' dan de modellen die aan LSTMs voorafgingen. Omdat er in natuurlijke taal vaak verbanden zijn tussen woorden die ver uit elkaar staan, zijn LSTMs heel geschikt voor NLP-toepassingen.

[41] Elk natuurlijke taal-analyserend algoritme heeft profijt van een zo informatief en rijk mogelijke representatie van de combinatoriek (context) van woorden.

BERT-modellen zijn relatief recent (2019) ontwikkeld door Google en worden sinds die tijd beschouwd als state-of-the-art taalmodellen. Deze modellen leren woorden te voorzien van een vector-representatie die allerlei contextuele informatie bevat: de woorden waar een bepaald woord doorgaans mee combineert. Zo'n representatie is uiterst informatief voor machine learning-gebaseerde (NLP, *natural language processing*) vervolganalyses. BERT-modellen worden getraind door bepaalde woorden te maskeren en te voorspellen welk woord er op de plek van het gemaskeerde woord zou moeten staan. Een tweede taak is gericht op het voorspellen of twee opeenvolgende zinnen een logische combinatie vormen. Als bijproduct van deze -tegelijk aangeleerde- taken leert BERT de toekenning van vector-representaties aan losse woorden. BERT heeft, anders dan eerdere modellen die van links naar rechts 'lazen', toegang heeft tot de context zowel links als rechts van een gemaskeerd woord - het is een bi-directioneel model. Varianten van BERT zoals RoBERTa richten zich uitsluitend op het maskeringsprobleem, of zijn aanzienlijk kleiner en daarmee sneller te trainen (DistilBERT).

Evaluatie Machine learning-modellen worden regelmatig geëvalueerd met zogeheten *F-scores*. De F-score is gebaseerd op een zogeheten *harmonisch gemiddelde* van *precision* en *recall*, en neemt waarden aan tussen 0 (een onbruikbaar model) en 1 (een perfect model). Hierbij wordt uitgegaan van een 'gouden standaard': vaak door de mens geproduceerde referentiedata waarmee de voorspellingen van een model worden vergelijken. Precision meet hoeveel van de voorbeelden die het model heeft aangemerkt als positieve voorbeelden (bijvoorbeeld een bericht op sociale media waaruit radicalisering blijkt), ook daadwerkelijk positieve voorbeelden zijn. Als het model bijvoorbeeld van zes voorbeelden voorspelt dat ze positief zijn, maar in werkelijkheid zijn maar vier van die voorbeelden positief, is de *precision* 4/6, dus 2/3 of afgerond 0,67. Recall meet hoeveel van de positieve voorbeelden in de data worden gevonden door het model. Als er in de data tien positieve voorbeelden zijn en het model vindt er vier, is de recall 4/10, dus 0,4. De F-score is, zoals gezegd, een soort gemiddelde van deze twee scores, en wordt berekend als *2 * (precision * recall) / (precision + recall)*. Bij de scores van hierboven zou de F-score dus 2*((2/3)*0,4)/((2/3)+0,4) = 0,5 zijn.

Er zijn twee redenen om in de literatuur gerapporteerde scores te relativeren. Ten eerste kunnen scores uit verschillende onderzoeken niet eenvoudig met elkaar worden vergeleken, omdat elk onderzoek andere datasets gebruikt. Het is dus niet bekend welke methode het best zou werken als je alle methodes zou toetsen met dezelfde data. Ten tweede bewijst een methode zich in de praktijk pas echt bij toepassing op echte data in plaats van op alleen de dataset die in een onderzoek wordt gebruikt.

Adversarial machine learning Machine learning-gebaseerde detectiemodellen kunnen om de tuin worden geleid door zogenaamde *adversarial attacks*. Dit zijn aanvallen op modellen waarbij subtiele aanpassingen in input data leiden tot misclassificaties. Zo kunnen modellen die gebruikt worden voor het analyseren van beeld in de war worden gebracht door een paar pixels in een beeld op een dusdanige manier te wijzigen dat het model de inhoud van het beeld anders interpreteert. Bepaalde beeldelementen -zoals wapens- kunnen zo zelfs aan het zicht worden onttrokken (zie Den Hollander *et al.*, 2020). Een mens ziet hierbij geen verschil tussen het originele beeld en het aangepaste beeld.

Voor het uitvoeren van een succesvolle *adversarial attack* hebben kwaadwilligen twee opties.. Een zogeheten *white box attack* bestaat uit het verwerven van een *volledige* kopie van het getrainde bronmodel, met al zijn technische details, parameterinstellingen, en datarepresentatie. Een *black box attack* bestaat uit het stapsgewijs afleiden van een

gedeeltelijke kopie van het bronmodel, door bijvoorbeeld systematisch data (*queries*) te sturen naar een AI model (dat bijvoorbeeld beschikbaar is als een online service, zoals ChatGPT of Midjourney), en de uitvoer (zoals beeldinterpretaties) te verzamelen. Vervolgens kan er met diverse methoden onderzocht worden waar de kwetsbare plekken in het kopiemodel zitten. Voor beide typen aanvallen geldt dat een succesvolle ondermijningsstrategie gericht op de kopie gegarandeerd toepasbaar is op het bronmodel, hetgeen een verrassend resultaat is voor *black box attacks*.

Bijlage B
Workshop 2 resultaten

Hieronder vermelden we de woordelijk getranscribeerde, geanonimiseerde responses van de Workshop 2-partners op een aantal gepresenteerde stellingen. Vanwege het internationale karakter van de deelnemers was de voertaal Engels.

Thesis 1	**Total Yes**	**Total No**
We are fully aware of the new threats emerging from AI and we can detect and handle these adequately.	4	10
Answers	**Yes**	**No**
No, we are not sufficiently trained on the threats of AI. We especially lack knowledge on tooling that may be helpful to detect threats.		X
We are aware, are working on detecting and handling. This is the nature of evolving fields.	X	
There is still a lot to explore and as usual, malignant actors will find creative reasons to exploit and circumvent.		X
We are getting a good sense but it's extremely early.		X
We are aware but haven't dealt with it yet.	X	X
These new technologies will bring abuse we cannot imagine in fields we cannot think of now.		X
Implications are not clear yet, so we're definitely not ready for them.		X
We are still learning about the threats and how, if at all, we can/should respond.		X
We seem to be aware of the potential dangers, but seem not able to deal with them at this moment.	X	X
We might describe what AI has produced thus far, yet we can't oversee impact in the field.		X
Not fully aware of how it will be misused. Technology to tackle synthetic content exists, but is not refined/mature enough.		X
We spend a lot of time revising policies and red teaming.	X	

Thesis 2	Total Yes	Total No
Our staff is continuously monitoring the scientific aspects of extremism detection.	7	5
Answers	**Yes**	**No**
We share key pieces of research among the team but also use the data we collect to generate findings.	X	
We have dedicated units looking at innovation and specific platforms bridging public-private exchanges.	X	
Our focus is on terroristic content. We do monitor scientific aspects but not as a primary activity.		X
We monitor the latest developments passively, but don't have the time to look for them in a proactive manner.		X
Our department has a department dedicated to this.	X	
Not following new trends. Only reading popular media.		X
We are not equipped to do that. Almost no hosting/cloud provider has these resources.		X
We look at the practical side, like consequences.		X
Yes, but still in an early phase of development, so more needs to be done and incorporated into strategies.	X	
Yes	X	
We monitor and analyze academic journals and apply relevant scientific/academic input to our approach.	X	
We have dedicated teams reading up on new research and connecting with peers.	X	

Thesis 3	Total Yes	Total No
We favor detection methods that are highly accurate but may miss some things, as opposed to methods that are less accurate but capture a lot (and demand humans to check their results).	7	5
Answers	**Yes**	**No**
We prefer to find the sweet spot but from a Counterterrorism perspective we can't afford (automated) prioritization errors.		X
Our purpose is to detect material to prevent it [from] going viral, and clean up the web in cooperation with tech. All material needs to be checked by a human and we want to know the status quo of material online.		X
As law enforcement we cannot afford to miss certain elements.		X
We do targeted OSINT collection based on specific criteria and feed into TCAP [Terrorist Content Analytics Platform] platforms.	X	
We use wide breadth analysis for detection but we use high precision for action.	X	X
Both, depending on harm and remedies.	X	X
We use systems that flag potentially violative content. We use systems that flag useful signals, but some noise is inevitable.		
We do not have resources to do massive manual classification.	X	
Yes, and manage capacity.	X	
No clear opinion, but from th perspective of human capacity, it should remain workable.		
High accuracy: reduce risk of stigmatization; human assessment may still be required.	X	
Accuracy, focus on clearcut reliable signs, although the volume might be smaller. Matter of priorities.	X	

Thesis 4	Total Yes	Total No
Explainability of detection methods is very important.	11	1
Answers	**Yes**	**No**
Transparency is important for us – from OSINT to TCAP collection	X	
If we cannot explain how we find evidence, it will not be used in court and could only serve intelligence purposes.	X	
Since all of the identified content is forwarded to other stakeholders. Sources and metadata is important to share too.	X	
We need to detect and send out a removal order.		X
Yes, for policy and legal reasons.	X	
Yes, explainable to the users.	X	
For accountability and transparency, but also for keeping control and learning of what you are doing.	X	
Both the analysist and user should know why content is flagged/considered undesirable and leads to behavioral interventions.	X	
Our information might inform legal cases and it's validity should be transparent from an early stage.	X	
We test systems and document how they work.	X	
We have seen algorithms going wrong. Legal certainty is very important.	X	
Decisions should be explainable and documented if they have consequences.	X	

Thesis 5	Total Yes	Total No
Detection tools always need to have humans in the loop. They must be interactive.	6	5
Answers	**Yes**	**No**
I don't think there is added value for these tools at this level, nothing that could be done by humans better.		X
A dialogue-level detection took isn't necessary, but humans must look at the outcome.		X
For law enforcement purposes human eyes are crucial to verify detection due to consequences.	X	
Human review is important – the two should inform each other.	X	
Yes, but this is very resource heavy.	X	
Depends on precision of the tool, scale and consequences of the tool outcome.		X
All seems inadequate for this moment.	X	
From a legal perspective, a human always has to make the definitive decision. You want to adjust your tooling when you find the algorithm is not right.	X	
There needs to be the possibility of human intervention.	X	
It depends – certain tools can take down certain content with very high accuracy. For radicalization, context and human in the loop are needed.		
In EU LSAM cases human eyes are not necessary.		X
No		X

Thesis 6	Total Yes	Total No
We are hampered in our detection work by Societal, Ethical, Legal and Privacy principles imposed by law and other (EU, national) regulations.	6	2
Answers	**Yes**	**No**
Especially because of the rapid developments of tech/AI and the slow development of legal frameworks.	X	
All actions should be grounded by rule of law and right to privacy.		X
We don't have SELP principles enough.		X
Intelligence sharing is suffering from wrongful explanation of law/regulation.	X	
Detection can only be biased.	X	
"Massive" data collection is restricted for multiple reasons: legal, societal, ethical.	X	
Unsure; our possibilities are limited, but if information sharing was streamlined, this could be overcome.	-	-
Not sure whether regulations are an obstacle. First problem is bureaucracy.	-	-
Legal, direct messages, content and metadata, facebook data sharing	X	
Data protection limits the detection work we can do. There are also ethical/moral/philosophical concerns.	X	

Distributielijst

ONGERUBRICEERD Releasable to the public ⟩ TNO 2024 R10310

JENV

Programmabegeleider NCTV M.T. Croes, Kenniscoördinator	pdf
Programma referent NCTV M. Kowalski, Decaan NCTV Academie	pdf
Projectbegeleider NCTV D.J. Sent, Beleidsmedewerker CT Online	pdf
Directie X	
- x@minjenv.nl	pdf
- H. Hanoeman	pdf
- B. ter Luun	pdf

TNO

Ten behoeve van opname in bibliotheek en distributie binnen TNO door:

TNO Bibliotheek (RIS) locatie Soesterberg	*link*
VP-manager VM/KOP G.R. Jansen-Ferdinandus	email-alert
Projectleider J. De Schepper	email-alert
Research manager projectleider A. Woering	email-alert
M. van Berlo	email-alert
E. Poot	email-alert
A. de Vries	email-alert
D. van Hemert	email-alert
S. Raaijmakers	email-alert
D. Blok	email-alert

Defence, Safety & Security

Oude Waalsdorperweg 63
2597 AK Den Haag
www.tno.nl

TNO innovation for life