ISD | Powering solutions to extremism, hate and disinformation

POSITIVE ONLINE INTERVENTIONS PLAYBOOK:

Innovating Responses to a Shifting Online Extremist Landscape in New Zealand

This playbook was developed with the support of the New Zealand Department for Prime Minister and Cabinet and the Department of Internal Affairs.

It contains images that may be distressing or cause offence.

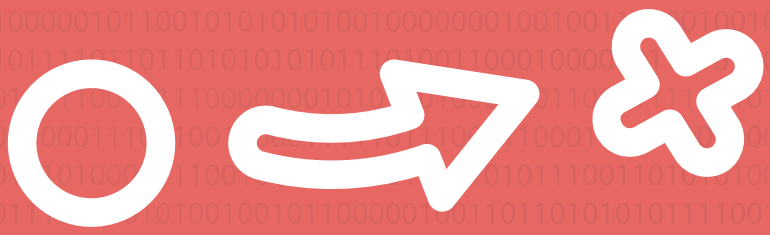
Amman | Berlin London | Paris | Washington DC

www.isdglobal.org

Over the last decade, the international extremist threat landscape has transformed dramatically, as a threat largely characterised by recognisably violent groups shifts towards more amorphous online extremist ecosystems. This nebulous landscape has facilitated the transnational spread of extremist ideas and strategies with highly local impacts, as the lines between diverse extremist communities, conspiracy movements and hate groups are increasingly blurred.

A suite of highly securitised responses developed in the aftermath of 9/11 - often narrowly targeted at Muslim communities - are not fit for purpose in responding to this transformed threat. In this increasingly hybridised threat environment, new proactive approaches are urgently needed which are rooted in a public health approach, which target the causes rather than symptoms of violence. Community-led designs are needed that respond to the evolving shape of extremism, and which bridge online and offline engagement in a context where these distinctions are increasingly arbitrary.

This playbook considers the implications of these profound shifts for positive online interventions efforts, including programming aimed at building digital literacy, communicating with key online audiences, and proactively engaging with those at risk online.



It takes stock of existing approaches, brings together domestic and international best practices, and suggests potential avenues for developing new positive intervention strategies to counter extremism in Aotearoa New Zealand. Beyond narrow efforts to address violent extremism, the models and approaches to intervention in this playbook aim to build resilience to a range of a range of societal harms, including other forms of violence and online exploitation.

This playbook has been developed in consultation with 40 New Zealand civil society organisations and communities. The authors are grateful for the many organisations who contributed time and insights, particularly Māori and Pasifika communities. Recognising that prevention must be rooted in local community, this playbook is intended as a framework for adaptation and delivery by the rich spectrum of civil society groups, practitioners and communities working to address this constellation of challenges in a New Zealand context.

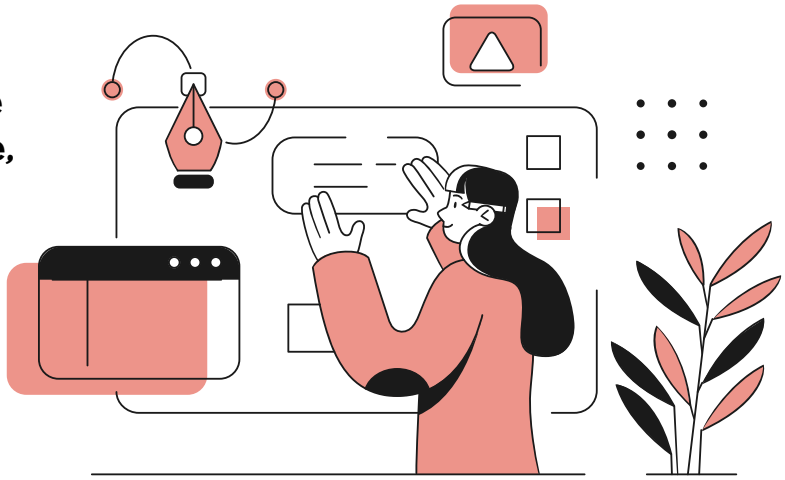
To inform evidence-based approaches, the playbook starts by outlining six key trends in the online extremist landscape, examples of their local manifestations and their implications for intervention. Secondly, it outlines the strategies, successes and challenges of existing intervention models, as well as emerging and innovative approaches. Finally, the playbook reflects on practical considerations needed for civil society to pursue this work, including monitoring and evaluation, safeguarding, operational security and ethical considerations.

Contents

CONTEXT

Trends in the Evolving Threat Landscape

The extremist social media landscape is a vast, rapidly evolving ecosystem comprised of inter-connected groups, networks and individuals. This section captures some of the key trends in the contemporary international landscape, their manifestations in New Zealand and how they impact on intervention approaches.



Ideological Hybridisation

Blurry lines between extremist ideology, hate & disinformation

Organisational Decentralisation

Increasingly diffuse 'post-organisational' threat environment

Transnationalism

Cross-border extremist organising and influence

Aesthetics & Culture

The role of visual sub-cultures in socialising online extremism

Platform Amplification

How extremists take advantage of platform design features

Mainstreaming Extremism

Mainstreaming of extremist ideas through media and politics

Ideological Hybridisation

Trend

Hybridisation occurs as different online communities connect the dots across diverse, sometimes contradictory, ideological positions. Extremist ideologies used to derive from top-down groups and distinct literature; hybridisation has resulted in an increasing overlap and influence of different online extremist ecosystems, adjacent subcultures and hostile state actor activity.

For example, Figure 1, shared in neo-Nazi Telegram channels during the May 2021 Israel-Gaza war, depicts far-right extremist and Islamist networks uniting around a shared perceived 'struggle' against Jewish people. While this does not represent any actual collaboration, nor a suspension of virulent anti-Muslim hatred, it is one example of how extreme fringes amplify each other's ideas and content.

Figure 1: Image found on a white supremacist Telegram channel during the May 2021 Israel-Gaza conflict

The role of ideology in extremist radicalisation pathways has also evolved, where it can often be adopted as a final step by those with significant psycho-social vulnerabilities. Violent misogynists (including involuntary celibates 'incels') and school shooter-style plotters mimic the aesthetics and actions of violent extremist attackers and become mobilised in parallel online ecosystems. Such violence can have no clear ideological agenda but is adjacent to online extremism and is often still situated in a shared ideological backdrop of structural racism, misogyny and hate towards minority groups.

New Zealand Lens

New Zealand's conspiracy landscape is characterised by a tangled web of organisations and individuals across different online platforms. These may also subscribe to extremist ideologies such as white nationalism, but for many conspiracy theories are their sole motivation.

The mobilising potential of online conspiracies for offline violence was seen in extreme elements of the anti-lockdown movement, including at the 2022 Wellington protests.

The January 2022 car attack in Auckland motivated by extreme misogyny shows the blurring of ideologies. The attacker was not an overt 'incel', but his motivations included a grievance for his failure to have a girlfriend and previous fantasies of killing. The attack can be seen as an extreme part of a broader 'manosphere' which produces multiple other harms including inter-personal violence, domestic abuse, coordinated harassment campaigns, threats to reproductive health, exacerbation of structural gender inequality, threats to private and public safety, and threats to democracy.

Intervention Implications

Many efforts to tackle extremism are centred on ideological disengagement, which may prove futile where ideology is not central to engagement pathways. Other risk factors may be more relevant for intervention efforts, including social isolation, digital illiteracy and poor mental health. These factors have been particularly pronounced in young people who spend significant time online, with social psychologists drawing links between social media usage and loneliness. Various groups seek to exploit these vulnerabilities, including gangs, while misogyny is a common factor across different forms of extremism.

Opportunities therefore exist to integrate counter-extremism work within broader violence prevention strategies, such as efforts to tackle domestic violence. The public health approach outlined in the section below looks to address underlying vulnerabilities from a non-securitised perspective.

Organisational Decentralisation

Trend

While violent extremism previously mainly centred on formalised groups, the increasingly 'post-organisational' threat landscape is comprised of highly-networked but loosely-affiliated digital ecosystems. This is typified by the far-right mobilisation around 'leaderless resistance', which emphasises the ability of grassroots activists to take coordinated action.

The Centre for Research on Extremism has demonstrated that since the mid-2000s, 'far-right violence [in New Zealand] has been overwhelmingly perpetrated by lone actors or small unorganised groups'.

This trend was supported by the New Zealand government's recent counter-terrorism strategy claiming that "in New Zealand, if a terrorist attack happens over the next 12 months, it will likely be carried out by a lone actor."

New Zealand Lens

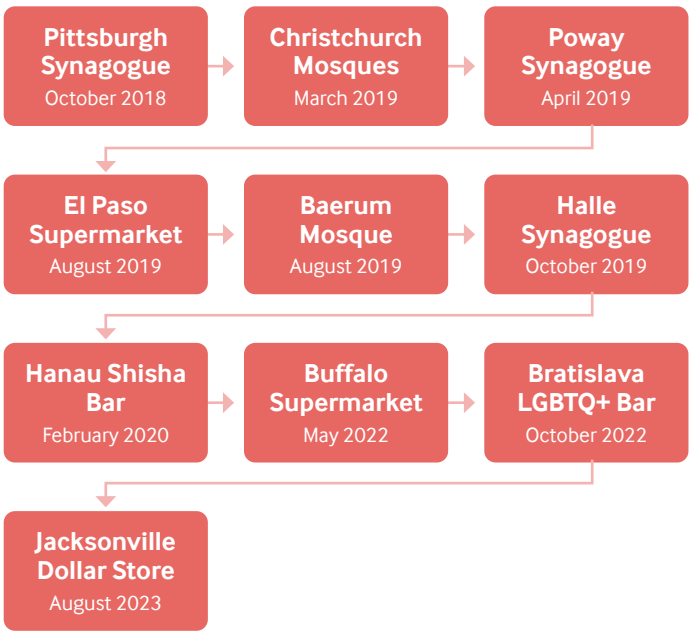
Closed, anonymous or pseudonymous networks including Chan forums, Discord servers and Telegram groups host markedly violent sub-cultures. Many are defined by their cultural codes, competitive vulgarity, and casual violent and hateful language. In New Zealand, organised extremist groups comprise a relatively small proportion of the extremist threat picture.

The 15 March attacks demonstrated the ability of an individual to radicalise and plan an attack in an online space without group backing. Its influence on future attacks including in Poway, California; El Paso, Texas; Halle, Germany; Baerum, Norway; Buffalo, New York; and Bratislava, Slovakia shows the reach of decentralised influences.

Intervention Implications

Approaches that have historically focused on curbing engagement aimed at groups with a clear ideological programme must adapt to a landscape of de-centralised networks and self-initiated individuals.

The Classification Office's designation of 'objectionable content' for propaganda from looser extremist networks – such as literature associated with the Neo-Nazi Terrorgram movement – shows the need to move beyond thinking solely about groups when understanding evolving online extremism threats.



Transnationalism

Trend

The global nature of social media has served to encourage the 'transnationalisation' of the extremist landscape, creating increasingly borderless threats.

This has multiple manifestations:

- The adoption of overseas narratives,
- Learning from international tactics,
- The formation of transnational networks with no particular host country.

Figure 2: Online mentions of international entities by NZ-based extremists (ISD 2021).

Increasing emphasis is placed on commonalities and shared identities (e.g. White identity), broadening the borders of more narrowly-focused conceptions of nationalism. For example, instead of the narrative of 'New Zealand for the New Zealanders', framings have pivoted towards the wider conception of 'European lands for European peoples'.

New Zealand Lens

Networks of extremist Telegram groups provide anonymous users, who may or may not be based in New Zealand, spaces to socialise.

Extreme-right wing New Zealand based group Action Zealandia has situated itself as part of a transnational network. It has engaged with this network through podcast appearances and activity on Telegram with groups such as US-based Patriot Front and the Nordic Resistance Movement. In their ethnographic research, Chris Wilson and James Halpin identify that in Action Zealandia online forums, "members are far more engaged in discussing international rather than New Zealand-related current affairs".

Intervention Implications

International coordination between governments, civil society and platforms are crucial for transnational responses in an increasingly borderless threat landscape.

Mechanisms like The Christchurch Call – a community of more than 130 governments online service providers and civil society organisations - work collaboratively and across borders to address the proliferation of violent extremist content.

Aesthetics and Culture

Trend

Figure 3: An example of 'fashwave' content.

In the online space, hateful ecosystems distinguish themselves based on aesthetic elements. These include memes, online gaming references and other visual reference points.

The sub-cultures formed online become areas for in-group socialisation, with users replicating aesthetics in their own content. Certain seemingly innocuous trends (such as 'Cottage Core') can serve as a pathway to entering Tradwife and other more harmful misogynist communities.

Visual and audio can be highly graphic, violent or offensive, and can appeal to 'thrill-seekers' or those seeking a community. ISD's research into the online 'Islamogram' phenomenon showed the merging of online Islamist communities with alt-right memes and gaming subcultures.

Figure 4: A meme showing the March 15 attacker being 'canonised' by Adolf Hitler and the Charleston church shooter.

New Zealand Lens

On extremist online forums, the 15 March attacker is revered as a 'saint' in a wider culture which deifies significant extreme-right terrorists. The attack serves as a rallying cry for future far-right extremist violence.

Popular gaming platforms including Minecraft, Fortnite and Roblox give players a great degree of freedom to construct their own worlds. This has resulted in, for example, replicas of the Christchurch attacker's gun or the al-Noor mosque being built; these are then interacted with by users of toxic online spaces.

Figure 5: The al-Noor Mosque has been built on Minecraft to glorify or digitally replicate the March 15 attacks.

Intervention Implications

Online interventions need to understand the significance of these visual dimensions within extremist sub-cultures. Attention should also be paid to the emergence of specific visual cultures as catalysts for hate and extremism, with the adoption of violent ideology often a subsequent step to involvement in harmful online communities.

Platform Amplification

Trend

Platforms themselves often surface harmful content to users who may not otherwise have seen it. 'Algorithmic amplification' occurs when platform algorithms, designed to maintain user engagement by recommending targeted content, promote harmful accounts and material to audiences.

Extremist networks increasingly understand content moderation policies and how to circumvent them. They achieve this through coded language (such as '88', a veiled reference to "Heil Hitler", or 'H0l0caust'), dogwhistles and other implicit hate speech.

Extremist networks use accounts on different platforms for different purposes. On mainstream platforms, where moderation can be stricter, they often post links designed to funnel users into more overtly hateful spaces.

ISD research on algorithmic amplification on YouTube Shorts showed how 10 new accounts registered as young Australian men were all organically served misogynistic and manosphere content by the platform's algorithm.

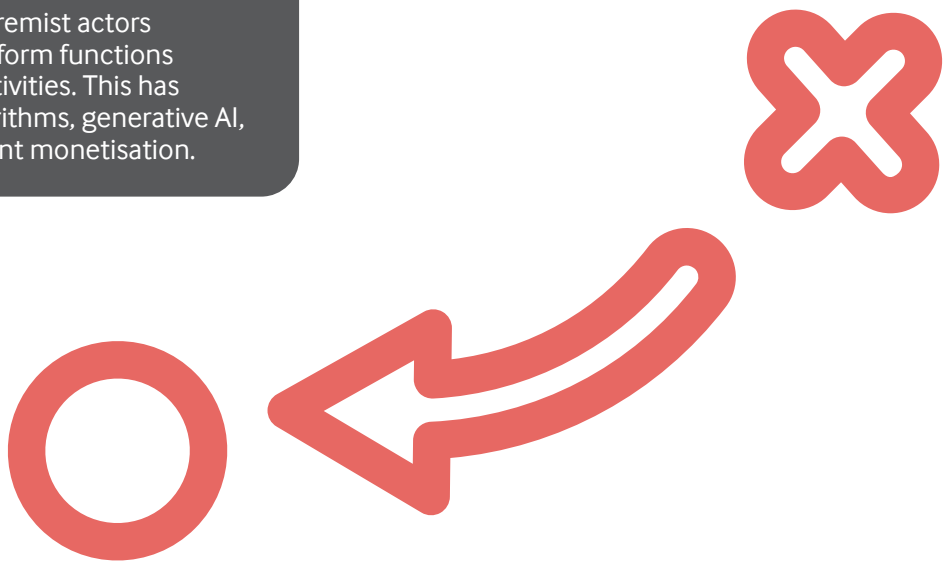
New Zealand Lens

In New Zealand, hate and extremist actors have taken advantage of platform functions to promote and monetise activities. This has included recommender algorithms, generative AI, augmented reality and content monetisation.

Intervention Implications

Whilst an important risk factor, algorithmic amplification and platform features must also be considered in the development of positive interventions. These include leveraging such systems for the dissemination of online campaigns.

Platforms themselves often offer programmes for amplifying counter-speech, including providing training and ad space for campaigns. Where platform features serve to promote hateful content, online interventions should go hand-in-hand with advocacy efforts to improve systems, policies and enforcement.



Mainstreaming Extremism

Trend

Extremist ideas do not just exist on the fringe of society but seek to gain acceptability in the mainstream and reach wider audiences. Their ultimate goal is to foster polarisation and undermine democracy.

The results of such mainstreaming go beyond violence: extremism erodes human rights, creating a chilling effect on civil society and further marginalising vulnerable communities. While extremists directly target communities with violence, this is often in parallel to more mainstreamed manifestations of hate such as transphobia.

It is therefore crucial to understand extremism's deep links to broader manifestations of structural racism, inequality and colonialism.

New Zealand Lens

Radicalised conspiracy movements have gained significant foothold in mainstream politics, sometimes boasting tens of thousands of members.

New research by the Classification Office highlights the long tail of harms of misogyny, from extremist male supremacism to the online abuse and harassment of women to intersections with domestic violence.

The Disinformation Project showed Māori and non-Māori women of colour to be the "primary targets of sustained, high-volume, networked targeting" on online platforms. Misogynoir is the unique denigration of women of colour, especially by white supremacists.

ISD research on the rise of QAnon in 2021 found the conspiracy rapidly shift from the fringe to mainstream in New Zealand during the COVID-19 pandemic; this culminated in protesters with QAnon flags clashing with police outside of parliament. New Zealanders sent more QAnon-related tweets per capita than the UK, Canada and Australia (1,500 Tweets per 100,000 Internet users), trailing only the US (the origin of the conspiracy theory).

Intervention Implications

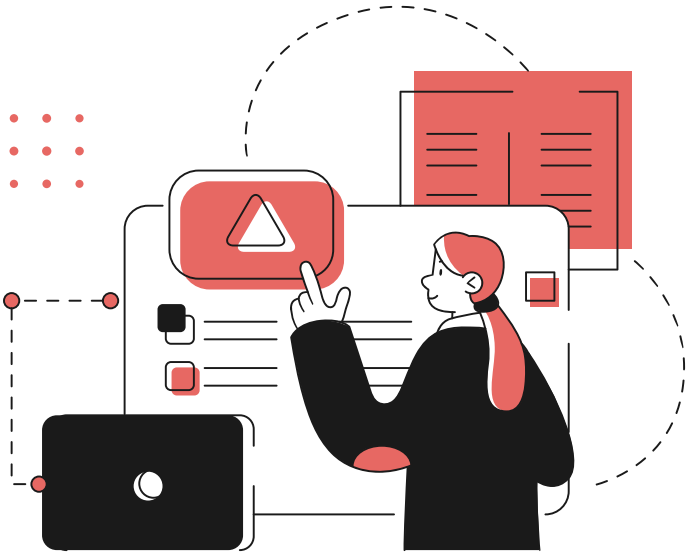
Interventions will need to address the growth of extremism across society, involving both the 'sharp tip' of violence and the slow creep of extremist narratives.

Such efforts will need to include a broader spectrum of responses than traditional narrow efforts focused on preventing violent extremism. These will include:

- Civic education and empowerment,
- Localised targeted support to communities,
- Democracy & social cohesion programming.

APPROACHES

Tried and Tested Intervention Models



Despite this dramatically evolving threat landscape, many approaches to (online) interventions have stayed the same.

Experts have warned that tried and tested approaches have largely focused on reactive rather than proactive tools including disruption, redirection and counter-narratives. This risks interventions that struggle to address contemporary threats and fail to recognise the emerging online subcultures and specific dynamics that define extremist movements today, such as extreme misogyny.

New proactive approaches are needed which are rooted in a public health approach. Community-led designs are needed that respond to the evolving shape of online extremism, requiring an understanding of the culture, aesthetics and narratives of different extremist groups and movements, as well as how they play off one another and increasingly converge.

From Reactive

Post-9/11 - responses to violent extremism developed, rooted in a highly **securitised approaches**, often disproportionally targeting Muslim communities.

Broad array of intervention programmes - e.g. one-on one support, disengagement and counter-narrative campaigns - designed for 'at risk' individuals.

To Proactive

Increased focus on **prevention**.

Adoption of a more **community-based approach** rooted in educational interventions focused on vulnerability, rather than simply deradicalisation.

Shift towards '**public health model**', targeting causes rather than symptoms of violent extremism.

Increasingly **bridging online and offline** approaches.

The Public Health Response

The public health model draws from wider harm prevention approaches; it seeks to build resilient communities and individuals by minimising 'risk factors' while boosting 'protective factors.'

These efforts are usually focused on four different 'tiers' of prevention:

- Primordial prevention works to foster healthy, resilient communities and individuals.
- Primary prevention builds community and individual resilience (to violence).
- Secondary prevention provides interventions for potentially vulnerable individuals.
- Tertiary prevention focuses on reaching individuals directly at risk (or even engaged in harm).

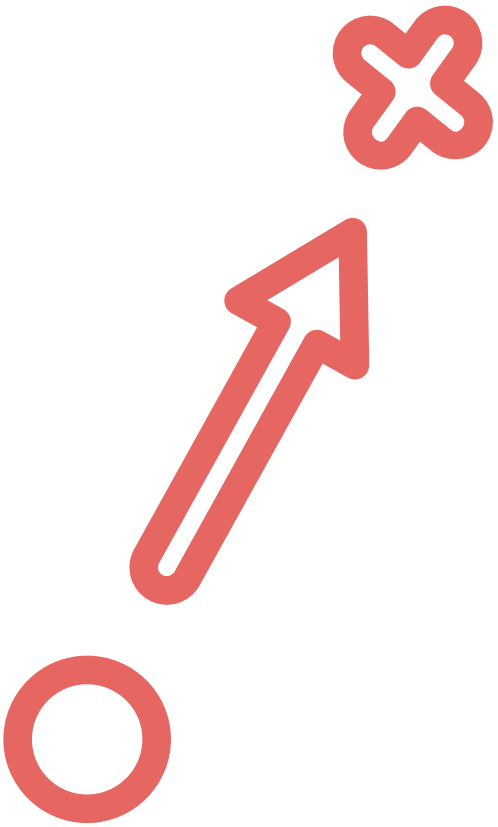
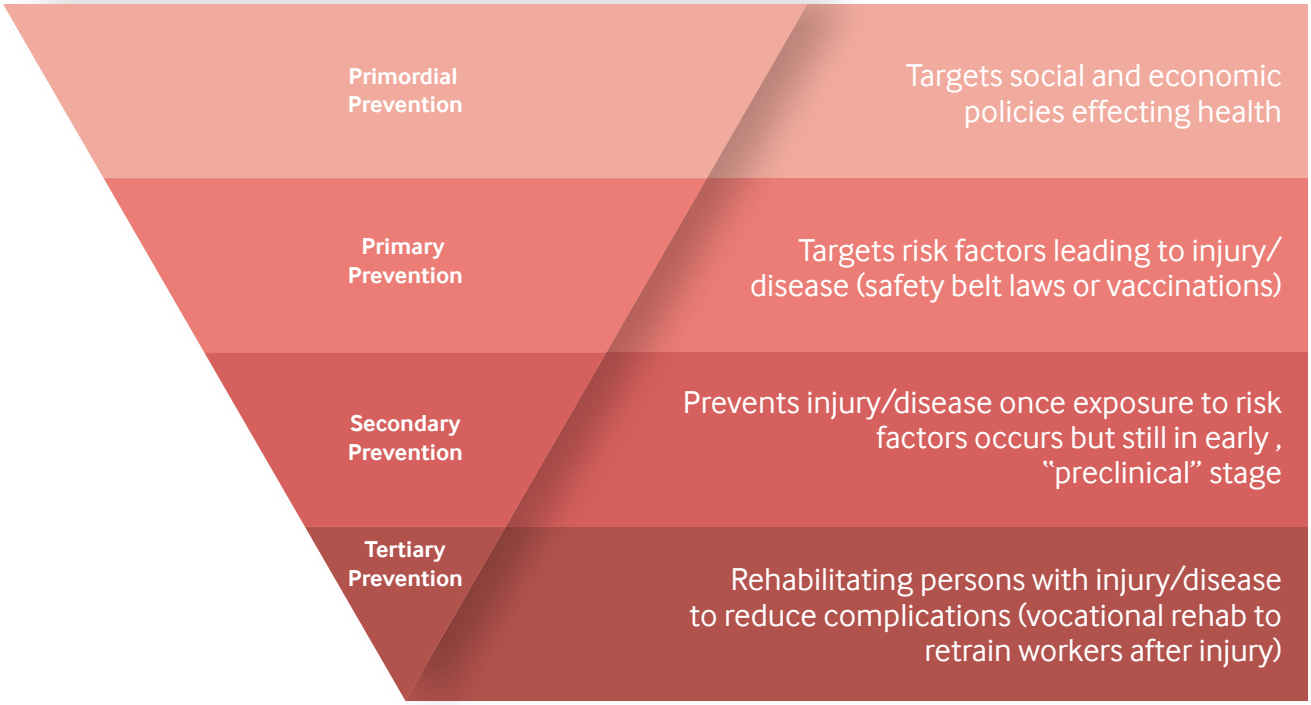
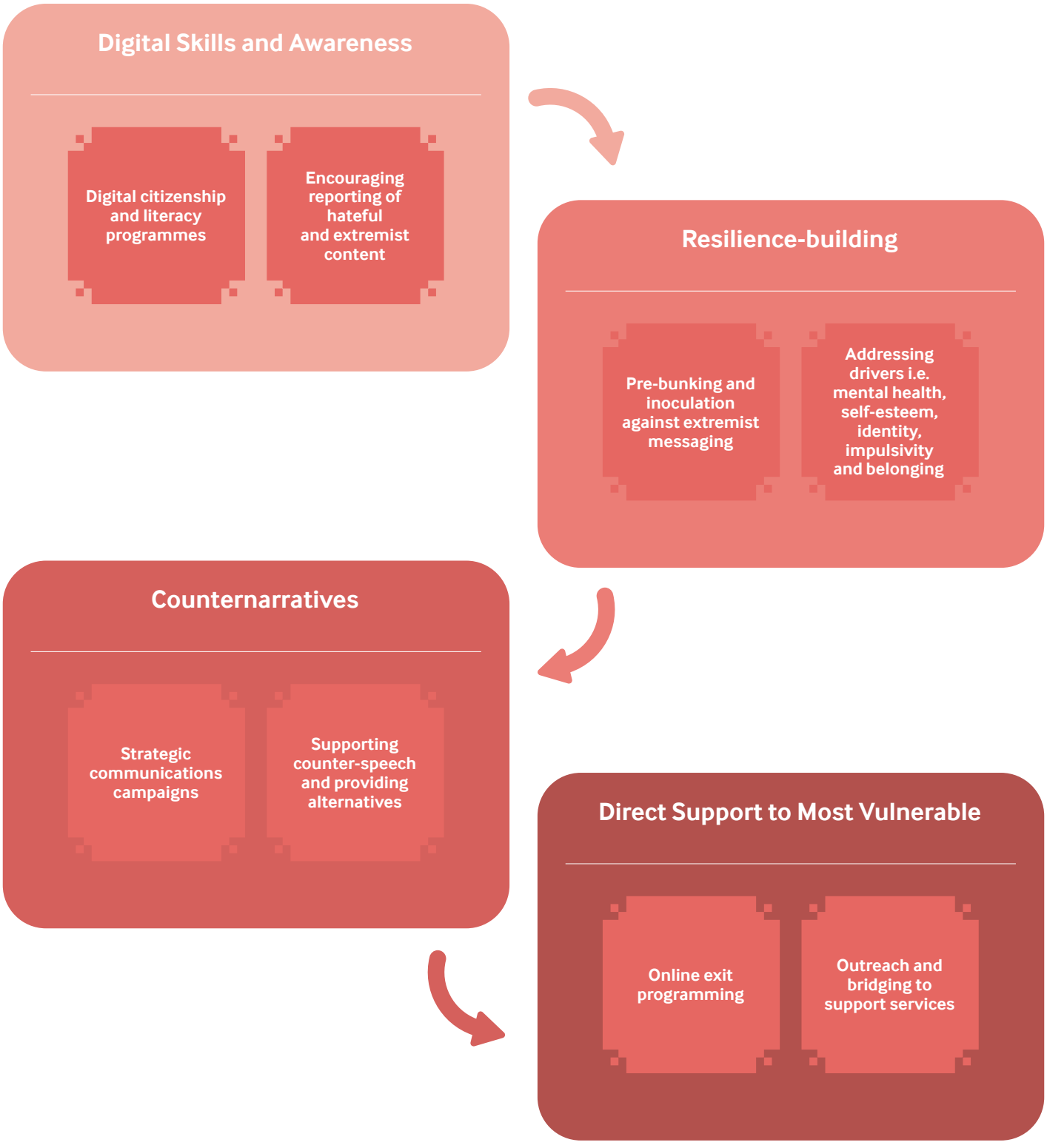


Figure 6 source: Based on tiered model of public health prevention, David Eisenman, 2016. National Academies of Sciences.



Tried and Tested Models

Based on the public health model, existing intervention models can be seen as ranging from 'upstream' to 'downstream', including the following categories of approaches:



Developing Digital Citizenship

Education in digital citizenship plays a crucial role in developing media and information literacy, advancing critical thinking, and enhancing understanding of how digital platforms function.

Whether delivered online or offline, these interventions can help people understand the harmful effects of online extremism, hate and disinformation; foster active participation online; and improve understanding of digital rights and responsibilities. Such strategies should not be seeking to teach people *what* to think, but *how* they can recognise harmful attitudes, behaviours and structures.

Online bystander training can help users report harmful content, support vulnerable communities and create a safer online environment.

Crucially, educational interventions should not stop with young people but include adults, as the impacts of online extremism cut across different generations in distinct ways.

Key topics for educational programmes:

- Identifying online mis- and disinformation, conspiracy theories, clickbait or manipulated media
- Understanding how polarising content can drive user engagement
- Algorithmic amplification and filter bubbles

CASE STUDIES

tohatoha

Tohatoha delivers evidence-based counter misinformation education to teachers and librarians, including an online 'escape room' game to help young people directly experience different misinformation tactics.

ISD's toolkit helps education practitioners develop and deliver effective Digital Citizenship Education, drawing on a decade's experience of youth programming, highlighting challenges and opportunities, and monitoring and evaluation approaches.

Countering Extremist Narratives

Over a decade of programming across platforms, ideologies, geographies and languages has shown a range of potential strategic communications opportunities for countering extremist narratives online:

- **Counternarratives** directly deconstruct or delegitimise extremist or hateful narratives, often through targeted engagement. These often target more "downstream" audiences who are either a part of or on the periphery of extremist groups.
- **Alternative narratives** offer a positive, empathetic, and hopeful alternative to extremist propaganda and may not directly engage with hateful narratives. These approaches often target more "upstream" audiences who are part of communities impacted by extremist recruitment and can include inoculation or "prebunking" to build resilience and a strong sense of identity. Alternative narratives can be *more impactful* than counter narratives, although they can be *harder to measure*. **Digital advertising techniques** used to target users based on perceived interests in extremist themes but can often be affected by rapidly shifting trends (memes, language), as well as an overreliance on specific keywords, hashtags, etc.

CASE STUDIES

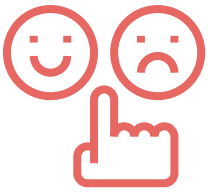
Tauiwi Tautoko mobilises and trains people to combat online anti-Māori racism, using models of allyship, community engagement and empathetic approaches.

The Campaign Toolkit provides concrete examples and frameworks for campaigns to combat online extremism, providing a range of strategies and resources for developing impactful approaches.

Challenges

Evaluating impact:
Evaluating impact has long proved challenging without deeper analysis of effects on long term attitudes and behaviour. It is therefore key that campaign objectives are specific, measurable and realistic.

Hostility:
Responding to hateful speech with hostility, aggression and insults can backfire, escalate, or deter others from joining any intervention.



Poor messenger:
Messengers trusted by target audiences are shown to have a more positive impact, while the wrong voice can further entrench views. Influential institutions and role models, such as football clubs, have shown promise in influencing hard-to-reach communities.

Unintentional amplification:
The US State Department's 2013 'Think Again Turn Away' campaign and similar projects show how approaches can backfire. In that case, it accidentally served to provide online ISIS supporters with a stage to voice their positions and engage in tit-for-tat fights with government accounts.

Direct Engagement: Support for Vulnerable Individuals

Direct digital engagement is a model for reaching vulnerable people, often applying tested offline intervention models in the online space.

This can include:

- Speaking to individuals via social media platforms and establishing an ongoing dialogue to promote alternative points of view,
- Seeking to off-ramp individuals into longer term support options, such as psycho-social care.

These interventions can be proactive, with intervention providers actively reaching out to individuals, or reactive, promoting the opportunity for discourse through advertising, media or social media posts. Existing community structures, such as Iwi or local systems of support, can serve as clear entry points for engaging trusted figures. These approaches have previously had huge success where communities are seen as partners not targets – for example, where Iwi were empowered to promote vaccine uptake among Māori communities.

Programming often incorporates prevention approaches, including mentoring, provision of psycho-social support or developing critical thinking skills.

Such approaches are highly complementary to programs like Just a Thought, which provides online Cognitive Behavioural Therapy to New Zealanders.

Challenges

- **Platform limitations:** Opportunities for reaching out to people are dependent on social media platform functionalities, and establishing a rapport with someone can take significant time and effort.
- **Offline support**: Having an appropriate offline support programme to refer vulnerable individuals to can be a challenge. It is essential for engagement-based programmes to have clear referral pathways built into their design.
- **Scepticism**: Online engagement is less effective in dealing with significantly radicalised individuals who are sceptical and less open to opposing viewpoints. However, online engagement can be a starting point for a more in-depth, sustained intervention.
- **Hostility and Disengagement**: Engagement-based interventions are not a silver bullet. Individuals may be hostile to outreach or suddenly disengage. It is essential that online engagement is professional, sensitive, and avoids antagonism.

CASE STUDIES

Counter Conversations: ISD's pilot program applied offline counter radicalisation interventions models online, directly engaging those at risk.

Intervention providers were former members of extremist movements, social workers and counsellors. They reached out over Facebook messenger to individuals expressing support for far-right extremist and Islamist movements.

They were able to have a sustained conversation with more than two-thirds of people they reached out to, and one-in-ten of these conversations demonstrated evidence of positively impacting on the candidate.

Violence Prevention Network (VPN): VPN translates an established offline intervention methodology – developed over two decades of engagement with extremists in Germany – into online spaces.

This includes conducting hybrid interventions with at-risk individuals and combining face-to-face communication and online engagement.

Ask an Aunty provides positive community-based spaces for young Māori people with the aim of combating disinformation and conspiracy theories. Rooted in community structures and principles of collectivity, it carefully considers the role of trusted figures in information dissemination.

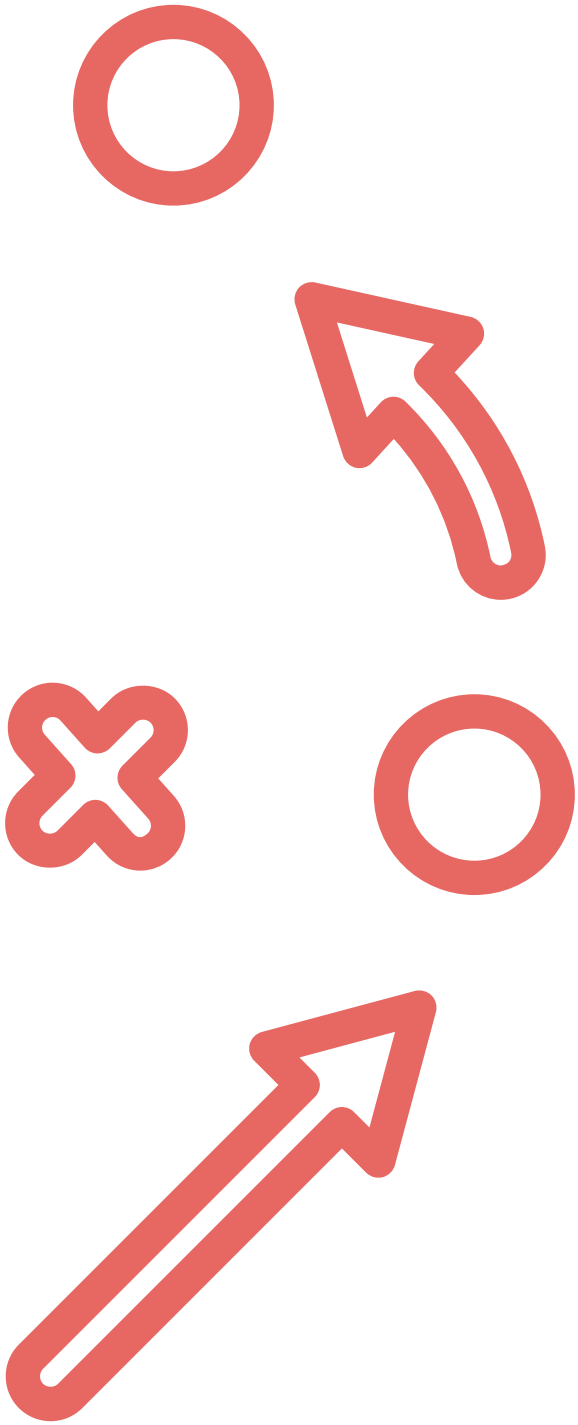
New Approaches for a Changing Threat Landscape

The tried and tested models explored above have largely been geared towards addressing specific manifestations of violent extremism.

The changing extremist landscape – with threats less explicitly group-based, more transnational and increasingly intersecting with other challenges like conspiracies and state actor threats – requires a new toolbox of responses, beyond these existing paradigms.

In this context, key questions for programme design include:

- **Intervention:** How can interventions work to address looser extremist movements and bring people out of broader ecosystems, rather than groups?
- **Prevention:** What are the education programmes and community initiatives required to raise awareness of and build resilience against more hybridised extremism threats?
- **Location:** What are the online spaces for impactful interventions? These can range from youth-centred platforms like TikTok, to gaming adjacent platforms like Discord, to more adversarial and harder-to-reach online spaces like 4chan.



Innovative and Emerging Intervention Models

Inoculate and Inform

Inoculation-based strategies for challenging extremism are gaining traction as proactive and impactful 'pre-bunking' approaches. Across a range of fields, organisations have been able to scientifically prove that it has a positive effect on audiences and at scale.

For example, such approaches have been successfully used by the OECD to help hesitant people overcome their fears about the COVID19 vaccine and to build resilience to health misinformation.

By pairing inoculation programmes with awareness campaigns or, in some cases, gamification, modes of delivery can immerse users in their own journeys of discovery.

Target: Audiences potentially vulnerable to harmful content online.

Approach: Exposing audiences to small doses of content related to extremism, hate or disinformation to explain their flawed reasoning or the ways in which they are manipulative.

Academic studies of disinformation define inoculation as an "approach to messaging that builds up resistance to persuasion, influence, or manipulation attempts by exposing people to weakened or diluted forms of the same arguments."

These approaches aim to:

1. Enhance critical thinking,
2. Increase discernment,
3. Boost confidence.

For example, one study gave 357 participants a 'weakened' form of extremist narrative and forewarned them that it was manipulative. Participants then read extremist propaganda. The findings showed "inoculation positively predicted reactance to extremist propaganda, reduc[ing] intention to support the extremist group".

A Two-Step Inoculation Process:

Step One: Forewarning

Forewarning includes a combination of giving the audience a heads-up they are likely to encounter harmful content online, and a "microdose" of what that content may be.

This should be a misleading claim, whether a quote, meme, or short video clip that exemplifies the false narrative or manipulation technique.

Step Two: Refutation

Point out what is wrong about the example provided. For example: what is misleading, how the audience may be manipulated and the goal of the manipulator.

Some forms of a refutation message could include "discrediting the messenger" or "disputing the methods", such as the use of emotional manipulation or misleading use of statistics.

This goal of 'technique-based' inoculation is to educate audiences about manipulation tactics.

Examples

PERIL – American University

A series of 30-second inoculation videos – trialling both narrative and fact-based messaging – tested the use of inoculation in more than 1,900 unvaccinated individuals.

'Inoculated' individuals showed significantly greater ability to recognise and identify rhetorical strategies used in misinformation.

Universities of Cambridge and Bristol/Google Jigsaw

A team of psychologists from UK universities of Cambridge and Bristol created 90-second clips designed to familiarise users with manipulation techniques.

The videos introduce concepts from the "misinformation playbook", illustrated with relatable examples from film and TV such as Family Guy or Star Wars (e.g. "Only a Sith deals in absolutes").

The 'Bad News Game'

Players take the role of a fake news creator, gaining as many followers as possible while building credibility, growing from an anonymous social media presence to running a fake news empire.

The game creators found initial evidence that people's ability to spot and resist misinformation improves after gameplay, irrespective of education, age or politics.

Correct and Counter

In the current digital environment, conspiracy theories and extremist narratives often originate within fringe online communities before jumping to mainstream media and politics. It is therefore crucial to correct and counter potentially harmful narratives as close to real time as possible before they get the chance to take root and get further amplified.

This is even more urgent in a context where harmful online communities deploy AI-generated images, videos or audio, resulting in an informational environment ever more saturated with misinformation.

Target: Broad online audiences with a focus on more 'mainstreamed' extremism or conspiracies.

Approach: Real time monitoring of emerging fringe narratives at risk of reaching the mainstream can feed into a 'situation room', informing the crucial efforts of fact-checking organisations and online intervention providers to help reach potentially vulnerable audiences.

Key Considerations

- Consider time bound opportunities where analysis and response might be particularly impactful, for example around sensitive periods like elections where online harms might be heightened. As well as being less open-ended, this can allow for greater pre-planning around sustainable approaches to both analysis and response.
- Recognise the burden real time response places on volunteers and staff, and closely consider how to sustainably staff a rapid response function to avoid burnout and protect wellbeing and safety.
- Consider the diverse coalitions of researchers, CSOs and practitioners required to analyse and respond to the full broad range of online threats, from data analysts to OSINT researchers, and fact checkers to community groups.
- Targeted communities impacted by online harms can be the most valuable source of insights on emerging threats which can't be identified by broad digital monitoring. Consider mechanisms (and incentives) for different community groups to systematically feed in insights on what they are seeing at the frontline.

Examples

FACT Aotearoa - Fight Against Conspiracy Theories

FACT Aotearoa is a grassroots group of activists working to tackle harmful conspiracy theories and disinformation in New Zealand. By exposing online threats, the group actively pre-bunks harmful narratives, educates users about online manipulation, and provides practical resources for politicians, journalists, as well as friends and family of those at risk of conspiracy theories.

Ichbinhier

Ichbinhier operates an action group of over 40,000 members to write factual and constructive comments on Facebook to counter derogatory and aggressive voices in comments sections.

Disrupt and Dislodge

Attempts at disruption have typically focused on taking down harmful online content and accounts espousing and spreading extremist narratives. The resulting dynamic has kept such interventions in a constant state of whack-a-mole, leading to a cyclical pattern of responses and counter-responses.

To shift this dynamic, more downstream intervention models can help to disrupt these communities, dislodge their membership and outcompete their appeal. This goes beyond simply countering extremist narratives and towards tackling the foundational 'system of meaning' underpinning such communities.

Part of this effort is to drive a wedge between dedicated and more casual adherents of online communities, building additional frictions for those less engaged.

Target: Online extremist ecosystems and communities

Approach: Targeting communities and individuals involved in creating and amplifying harmful content in fringe communities on social media platforms, websites and image boards. Disrupting these ecosystems and narratives at their source are likely to have an outsized impact in limiting the spread of their ideas to mainstream audiences. Mobilisation strategies might take inspiration from other grassroots efforts, such as the Student Volunteer Army's crisis response structure.

Key Considerations

- It is important to create an ecosystem around any disruption campaign. The full breadth of the internet should be used – including websites, stand-alone accounts and influencers – to create an ecosystem mirroring those created by extremists.
- Rather than relying on traditional polished communication products, crafting the illusion of more amateur content production can help create the look and feel of user-generated content which online subcultures thrive on.
- Integrating people with experience in target communities is key to not only understanding how to create and design content, but also how to target the content to effectively disrupt and dislodge a specific extremist or conspiracy community.
- The role of humour should be carefully considered. As a counter-tool it has been found to be "polarising", especially if it seen as trying too hard or making light of serious issues. As such, it should therefore be approached in an authentic manner.

Examples

North Atlantic Fellas Organization (NAFO)

A decentralised online phenomenon that grew during Russia's invasion of Ukraine through coordinated targeting of prominent pro-Kremlin figures online.

NAFO is made up of an 'online army' which makes heavy use of memes to outcompete and drive disinformation accounts off social media. They have also raised funds for Ukraine through the donation of custom-designed digital avatars.

Google Bombing: ISIS-Chan

ISIS-Chan was a mascot created by Japanese users of an imageboard to counter propaganda produced by the Islamic State (IS). The campaign made use of Google bombing, which aims to take over search results for specific terms to spread a message.

The tactic has been used to disrupt and breakdown communications in a variety of communities.

The 'Baltic Elves'

Volunteers developed a strategy to counter Russian disinformation.

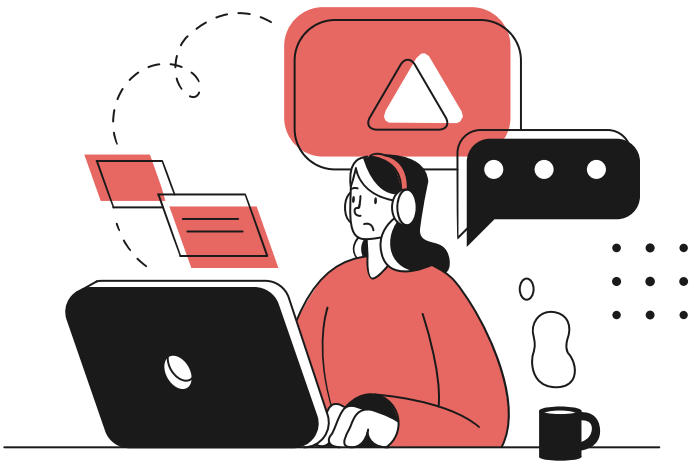
Operating across Europe, they monitor pro-Kremlin profiles and pages on social media, particularly on Facebook, and debunk disinformation through simple explanations and memes. The Elves' model is to get "the best people, the most committed people, and bring them all together in a much more collegiate and systematic way."

PRACTICE

Building a Model: Considerations for Developing Online Interventions

Monitoring and Evaluation

Before thinking about how to monitor and evaluate an intervention, it's crucial to have a **clear framing of the problem** the intervention is trying to address and a concrete understanding of the environment and context in which you will operate. This process takes a four-step approach, which can be used to develop approaches suitable for both long-term projects and reactive rapid response:



1 Situational Analysis

Developing a clear **situational analysis** while an intervention is still in its inception phase will provide a strong foundation for a design that can have lasting impact.

To effectively monitor and evaluate the impact of any intervention, it is crucial to first have a clear framing of the problem the intervention is trying to address.

A thorough situation analysis helps to ensure projects are tailored to the context:

- Identifying target beneficiaries which the intervention aims to reach and mapping potential partners that can help reach these audiences.
- Understanding the landscape in which the target audience operates, including important political or social factors underpinning the 'system of meaning' at the heart of extremist activity.
- Learning from past interventions, including their capacity gaps, opportunities and needs.
- Organising interviews and focus groups with key stakeholders and beneficiaries to understand their view on the proposed project, the problem it seeks to address and any environmental factors that may affect implementation.
- Determining what risks a project and its activities may have for its participants and developing strategies for mitigating these within a human rights framework.

2 Theory of Change

A theory of change helps define what success looks like for an intervention. Without them it can be difficult to know if an intervention has made a measurable difference to beneficiaries.

Goal	Ultimate long-term aim of intervention
Intermediate Outcomes	Medium-term results expected by end of implementation. E.g. Behaviour change.
Immediate Outcomes	Short-term effects on beneficiaries. E.g. changes in knowledge, skills, awareness, attitudes or access.
Outputs	The direct products and services delivered via activities. Outputs lead to outcomes but are not the change itself.
Activities	What an intervention *actually does*. These are actions through which inputs are turned into outputs.
Inputs	The human, financial, organisational and community resources required to implement an intervention.

Formulating measurable outcomes at the beginning of an intervention is critical if you hope to evaluate your efforts in the long run. Outcome statements clearly encapsulate the change you want to occur. It includes details on what will change, the direction of that change, who will experience it and where it will take place. For example:

Direction	What	Who	Where
Improved	awareness of reporting services for hate crimes	among minorities	in selected communities in city X
Increased	knowledge of how to advocate for human rights with the government	among civil society organisations	in country Y
Increased	ability to identify, critically engage with and resist extremist narratives	for at-risk youth	in region Z
Enhanced	equitable access to digital citizenship education programmes	for girls and boys	in deprived area X of country Y
Increased	perspective taking and tolerance of difference	by students	in schools in city X

3 Risk Assessment

Types of Risks

Reputational	Could the nature of the project be misconstrued or misrepresented? What effect could this have on those involved in the project?
Content	Is there a risk of intervention providers being exposed to harmful content due to the project? Can this exposure be mitigated?
Data	Is data being gathered or kept secure, especially any personally identifiable information? Are shared documents sufficiently protected?
Security	Could this work endanger the personal security of project staff, intervention candidates or the general public? What measures are in place to ensure their safety?
Legal	Is there legal risk associated with holding private information of individuals which could be subject to privacy requests?
Assumptions	Ensuring a strong evidence basis for the assumptions which underpin project design and delivery.

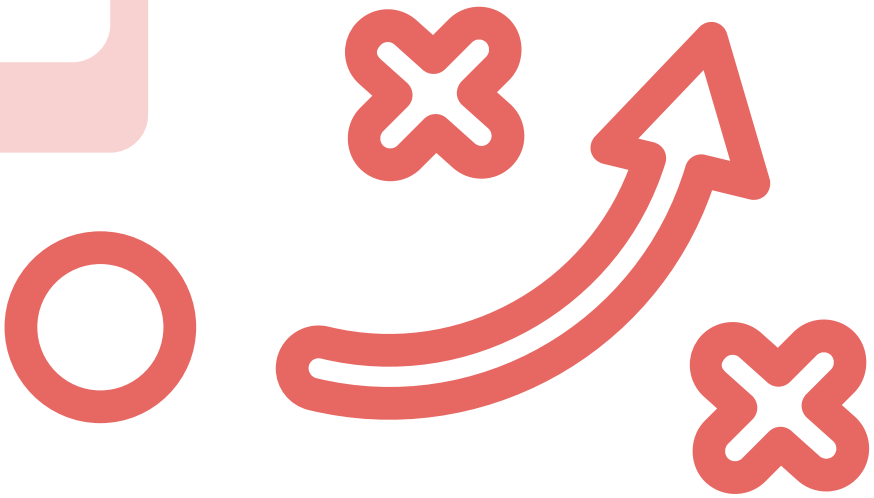
.

4 Data Collection

Depending on what is being measured, both qualitative and quantitative data can be collected.

Quantitative Data	Qualitative Data
Information expressed in numbers. Structured and rigid in its analysis. **Can answer:** "How much?" or "How many?".	Information expressed in words. More subjective in its analysis. **Can answer:** "Why?" or "What?"
Count: The total number of a given indicator (e.g. video views). **Rating scales:** The value a participant assigns on a rating scale.	**Perspectives**: What a person's views on a particular matter are. **Motivations:** Why a person behaves the way they do.

When developing a monitoring and evaluation (M&E) plan, several approaches to data collection can be used based on what is being measured, the nature of the intervention and the time available for analysis. These include observation, surveys, focus groups, interviews.



5 Evaluation

Various strategies can be employed to evaluate the overall impact of a programme on the toxicity of an online environment. These include:

Control/comparison groups: Leaving a group of participants without interventions. Comparing them against the group that did receive the intervention generates understanding of its efficacy.

Linguistic analysis: Studies suggest using qualitative and quantitative tools to comparatively analyse language on a platform before and after intervention. This can be challenged however by platform data access.

Message comparison: Assigning categories to counter-messages to facilitate relative comparisons. Whittaker and Elsayed suggest 5 categories: aggressive, argumentative, assertive, fact-checking and identity-building.

Panel surveys: Measuring the reach and impact of a campaign over different time periods recognising the non-lineal nature of change processes among target audiences.

Exposure analysis: Measuring the frequency that a particular campaign has been viewed by target audiences.

Contribution analysis: In recognition of the multiple concurrent programmes in any given field, this method understands the contribution of a project to the overall change.

Surveys: Using established survey instruments over individual questions can help holistically understand attitudinal change. Relevant examples of off-the-shelf evaluation surveys may include:

- Kirchner and Reuter's digital literacy scale measures respondents' intention to develop good practice in assessment of online mis and disinformation.
- The meaning in life questionnaire assesses respondents' sense of meaning and purpose in life.
- The general belongingness scale measures respondents' sense of belonging in their communities, including motivation to form communities and avoid isolation.
- The perspectives taking scale contains three measures to understand respondents' willingness to consider other people's viewpoints.
- The tolerance of difference scale contains eight measures to assess acceptance, respect and tolerance for diversity and difference in society.
- The Mayer-Salovey-Caruso Emotional Intelligence test (MSCEIT) measures the respondent's ability to "perceive, comprehend, act on, and manage emotional information".

Training measurements: The Kirpatrick model is a four-step process used to understand the efficacy of trainings and its link with change among participants.

Questions to Consider in Design Phase

- ❑ Who is the target of your intervention? Is it an individual or group? How has the intervention been tailored to best serve this target audience?
- ❑ What is the specific aim of your intervention? Is it to de-radicalise, prevent radicalisation or encourage disengagement from potentially radicalising groups/content?
- ❑ What is the scope of the intervention: nationwide, community-focused, individual?
- ❑ Will your intervention target identity, ideologies or mixed ideologies?
- ❑ What are the programme's components?
- ❑ Who are the key stakeholders involved in delivering this program? Why are these stakeholders best positioned to deliver the program?
- ❑ What would implementation look like? What logistical components should be considered? What resources? Scale of funding? What is the scope of human resources required (e.g. partner organisations)?
- ❑ What safeguards will be built in from the start to protect against unintended harms?
- ❑ How will success and progress be measured, with transparency and accountability at the forefront?
- ❑ How does this intervention program address existing gaps in intervention delivery in New Zealand? What is novel about it?
- ❑ How does the evaluation strategy consider the non-linear timeline of attitudinal/behavioural change? What longitudinal or qualitative impacts can reflect this?
- ❑ Does the project incorporate the necessarily extended timeline for evaluation?
- ❑ What level of evaluation does the project's resources permit?
- ❑ How does the evaluation strategy understand the specific contribution of the project in a busy sector?

Safeguarding

Both researching hate and extremism and working to counter these issues bring risks to safety and wellbeing. It is important to be clear-eyed on the potential risks to individuals working on these issues.

These risks could include:

- Virtual harassment or threats, including from extremists,
- Exposure to harmful and toxic content, including the potential for vicarious or secondary trauma responses,
- Physical threats.

These risks can be exacerbated for practitioners who identify as women, LGBTQ+ or are members of other targeted communities.

Strategies for reducing the impact of viewing harmful content (GNET)

- ❑ **Limit:** Limit time spent with harmful content and take regular breaks
- ❑ **Devices:** Limit viewing to specific devices
- ❑ **Audio:** Remove audio or reduce volume on video content
- ❑ **Visual:** Reduce screen brightness or apply a monochromatic filter
- ❑ **Text:** Use Optical Character Recognition to isolate text from visual content
- ❑ **Auto-play:** Turn off in platform settings
- ❑ **Blur:** Use an extension to obscure images
- ❑ **Support:** Ensure access to professional support to manage wellbeing
- ❑ **Community**: Create opportunities to engage with supportive peers

PEN America's guidelines for practicing counter-speech were crafted with a licensed mental health professional focused on wellness and safety. Below is their practical checklist:

- ❑ **Assess the threat level:** Conduct an honest assessment of both physical and digital security ahead of engagement.
- ❑ **Self-evaluate:** Am I ready for a confrontation? Engagement might escalate the ensuing abuse, which might have a more harmful impact on wellbeing.
- ❑ **Decide how and where you want to confront your harasser:** Consider whether it is better to address an individual, a community or a behaviour, directly or indirectly.
- ❑ **Establish your end goal:** Be clear on whether you want to accomplish awareness raising, correcting the record or changing minds.
- ❑ **Use language and craft messages that are likely to de-escalate the abuse:** Consider approaches rooted in empathy as well as condemnation, avoid hostile language, and root responses in concrete harms/consequences of online behaviour.

Operational Security

Where staff or volunteer's personal security is at risk of being accessed or publicly known about in harmful ecosystems, the Global Network on Extremism and Technology, VOX-Pol and Moonshot have recommended measures which can be taken to mitigate potential harm, including:

- ❑ Be aware about information available about you online and take steps to change privacy settings (e.g. havelbeenpwned, DeleteMe)
- ❑ Review both current and old accounts you may have forgotten about
- ❑ Use a reliable VPN (e.g. Private Internet Access)
- ❑ Set up a password locker (e.g. Bitwarden), randomised passwords and two-factor authentication
- ❑ Use a burner SIM card
- ❑ Use a virtual machine (e.g. Sandbox)
- ❑ Ensure shared files are hosted in secure systems
- ❑ Check your details are not publicly available on the electoral or vehicle registers

More guidance can be found in the Data Detox Kit.

Some systematic cyber security checks are available. For example, New Zealand-based Bastion Security Group's Help for Heroes programme provides pro bono cybersecurity support to not-for profits.

Reporting content

Where extremely harmful or potentially illegal content is seen, consider whether it should be reported to:

- ❑ Social media platforms – content which violates terms of service
- ❑ Classification Office – objectionable content
- ❑ DIA – online violent extremist material
- ❑ Police – 111 (emergency), 105 (non-emergency) including online form for threats.

Even where reporting or flagging mechanisms are not seen to be effective, their usage is important to prove both volumes of harmful content and potential failures of moderation.

Ethics

As well as security and security risks, there are several ethical risks which should be considered and mitigated against when designing interventions. From the 'do no harm' approach, interventions should be careful to understand when some more direct strategies, for example directly targeting extremist ecosystems, may raise specific ethical considerations.

Key ethical questions include:

- ❑ **The need to safeguard individuals engaging in interventions**
  How are you protecting participants from potential harms, including impacts on their wellbeing, employment and reputation? Are you ensuring their anonymity?
- ❑ **Consent**
  For targeted ongoing interventions, are recipients providing informed consent for engagement? How are you addressing possible power imbalances between you and recipients/ participants?
- ❑ **Unintended consequences of interventions:**
  How are you ensuring that your intervention does no harm? Are you considering the possibility that it might have unintended consequences? Are interventions appropriate for the target audience? Have you considered the risk posed by the target group?

Conclusions and Recommendations

Civil society and communities play a central role in the development and delivery of positive online interventions which cut across different levels of prevention, rooted in a public health and rights-centred approach.

While some existing efforts have been promising, others have not reached or effected attitudinal change in their intended audiences; some have even backfired. All serve as opportunities for further understanding success in interventions. Promising new approaches include efforts which look to inoculate, disrupt and counter the harmful impacts of online extremist in real-time. But such models need to be rooted in systematic evaluation, with consideration about concrete success metrics, risks and outcomes. At their heart, they should centre the safeguarding of intervention providers, operational security for project staff and volunteers, and a clear ethical framework guiding implementation.

This playbook has sought to offer a guide for the delivery of innovative intervention strategies. There are diverse approaches to interventions, and this playbook has attempted to highlight a non-exhaustive range of learnings from previous work. However, impact can only be achieved with the input on the ground of New Zealand's rich civil society sector, who bring extensive local knowledge, community buy-in and huge energy to the work.

But such efforts require meaningful support from government and other sustainable funding sources to ensure that such programming can be delivered in a strategic way. There is a particular need for policy makers to listen to community needs rather than approaching them transactionally as service providers. This also means engaging and sourcing ideas from a wide range of communities impacted by online harms, not just a smaller group of 'usual suspects'.

Challenging issues like extremism will impact on different communities in different ways, and some underrepresented communities may require particular support to help them understand their important role in delivering impactful programming. In particular, there may be specific work needed to translate the more technical dimensions of policy areas like 'P/CVE', language which can also serve to be divisive within some communities and sectors.

Given the transnational nature of the threat, there is furthermore a clear opening for greater international exchange between practitioners. A mechanism for ongoing collaboration and sharing of best practice across borders would help to bridge global problem sets with local solutions.

Finally, complementing the establishment of a more sustainable community of practice for prevention, there is also need for greater cross-governmental working with communities, which brings together corrections, education, international and domestically focused policy makers to address these challenges in a more coordinated manner.

Appendix 1: Additional Resources

Context

He Whenua Taurikura – Centre of Research Excellence

Te Mana Whakaatu – Classification Office

Hate & Extremism Insights Aotearoa

The Disinformation Project

'Digital Violent Extremism Transparency Report 2023', Te Tari Taiwhenua – Department of Internal Affairs, 2024.

'Understanding the New Zealand Online Extremist Ecosystem', ISD and CASM Technology, 2021.

'Online Misogyny and Violent Extremism: Understanding the Landscape', Classification Office, May 2024.

Approaches

Joe Whittaker et al. 'Unleashing the Potential of Short-Form Video: Strategic Communications for Countering Extremism in the Digital Age', Swansea University, RUSI, Hedayah, 2024.

'Campaign Toolkit', ISD.

'Content-Sharing Algorithms, Processes, and Positive Interventions Working Group: Part 2: Positive Interventions', GIFCT, 2021

Henry Tuck & Tanya Silverman, 'The Counter-Narrative Handbook', ISD, 2016

'Building your own Counter-Narrative Campaign on a Shoestring', DO One Brave Thing, 2020.

Nafees Hamid, 'Mass media and persuasion: Evidence-based lessons for strategic communications in CVE', XCEPT, 2022.

In Practice

'RAN guidelines for effective alternative and counter-narrative campaigns (GAMMA+)', RAN, 2017.

Tim Hulse and Michael J Williams, 'Shared Endeavour Fund Call Three Evaluation Report', Mayor of London Office for Policing and Crime & Strong Cities Network, 2024.

'Strategies for Researchers Analysing TVEC', GNET.

'Data Detox Kit'.

Elizabeth Pearson et al. 'Online Extremism and Terrorism Researchers' Security, Safety and Resilience: Findings from the Field', VOX Pol, 2023.

'Digital Security Resource Hub for Civil Society', Amnesty International.

ISD | Powering solutions to extremism, hate and disinformation

Amman | Berlin London | Paris | Washington DC

www.isdglobal.org