
GIFCT
Global Internet Forum to Counter Terrorism

# Content-Sharing Algorithms, Processes, and Positive Interventions Working Group

## Part 2: Positive Interventions

July 2021

GIFCT

Global Internet Forum to Counter Terrorism

## Executive Summary

**Over the past 12 months, representatives from government, tech, and civil society have come together as part of the GIFCT Content-Sharing Algorithms, Processes, and Positive Interventions (CAPPI) Working Group (WG). The group adopted the shared goal of mapping content-sharing algorithms and processes used by industry that could facilitate consumption of content that may increase user interest in or amplify terrorist and violent extremist content and consider positive interventions and risk mitigation points. This report focuses on the second part of that goal in mapping and considering positive interventions and risk mitigation points. This multi-stakeholder exercise seeks to consolidate examples of positive interventions from across sectors within a theoretical framework and inform the reader to increase awareness and understanding. Through these examples key findings and recommendations are provided. The GIFCT CAPPI WG members recognize the distinct strengths and capabilities of civil society, academia, government, practitioners, and the technology industry. Drawing in part on the findings, the recommendations are offered with a multi-stakeholder approach in mind.**

Working groups are a multi-stakeholder effort to further discussion on the given topic of the nexus between terrorism and technology. This paper represents a diverse array of expertise and analysis coming from tech, government, and civil society participants. It is not a statement of policy nor is this paper to be considered the official view of the stakeholders who provided inputs.

# Introduction

GIFCT Member companies prohibit terrorist and violent extremist content (TVEC) under their terms of service, community guidelines, or content policies. This is one prerequisite to GIFCT Membership and means that platforms remove this content if and when they become aware of it on their services. However, content removal is just one lever platforms may use as part of their broader content moderation efforts, particularly when it comes to tackling some of the not illegal yet potentially harmful drivers of extremism. In categorizing these other levers, they can be considered as positive interventions.

Following the Christchurch attack in New Zealand on March 15, 2019, several tech companies and governments joined the Christchurch Call to Action. In joining the call, industry members committed to "review the operation of algorithms that may amplify terrorist and violent extremist content." In July 2020, GIFCT established two WGs focused on algorithms and positive interventions to practically take forward the call commitment in a multi-stakeholder forum. These two WGs were then combined to form one single working group. The CAPPI working group is made up of representatives from governments, tech companies, and civil society, including academia, practitioners, human rights experts, researchers, and members of the NGO community.

In setting out the work program for the past 12 months, the shared objective of the group was to collaborate across industry, government, and civil society "to map content-sharing algorithms and processes used by industry that may facilitate consumption of content that may increase user interest in or amplify terrorist and violent extremist content and consider positive interventions and risk mitigation points."

This report focuses on the second part of the objective in considering positive interventions and risk mitigation points. GIFCT's mission is focused on preventing terrorists and violent extremists from exploiting digital platforms. For the purposes of this work we have taken a content-agnostic approach and sought to map positive interventions already used in the countering violent extremism (CVE) space as well as those that have been used elsewhere but could be applied to CVE. The use of counter and alternative narratives is recognized as an online CVE mechanism for connecting users to alternative sources of information. Yet there is less awareness of the uses of other forms of positive interventions.

This document maps and consolidates examples of positive interventions that have been used by industry, governments, and civil society practitioners in a theoretical framework to build increased awareness across the GIFCT multi-stakeholder community. It considers the full range of objectives, strategies, and tactics that are worthy of consideration by GIFCT members and highlights case studies of positive interventions that bring these components to life. Finally, in mapping these positive interventions, the document makes a series of multi-stakeholder recommendations for potential areas of future work.

This paper may also be read in conjunction with CAPPI Part 1: Content-Sharing Algorithms and Processes.

## Theoretical Framework

**Positive interventions** have a two-part definition that encompasses a What and a Why:

1. **What:** The promotion of credible, positive alternatives or counter narratives, and other forms of digitally distributed user-facing messaging;
2. **Why:** With the goal of counteracting the possible interest in terrorist and violent extremist groups.

Such a short definition gives a near endless set of possibilities that are not particularly useful. A theoretical framework helps GIFCT members to prioritize among these options and identify case studies relevant to the chosen approach. In time, such a framework will help the GIFCT community to group similar **positive interventions** together and measure their impact consistently. This will help the GIFCT community to learn from and iterate these approaches, allocate resources more appropriately, and increase impact.

## Preventing/Countering Violent Extremism (P/CVE) Objective of Positive Interventions

The first vector for this theoretical framework is the P/CVE objective the positive intervention seeks to achieve. This is inextricably tied to the audience, as defined by their relation to the act of violent extremism, and can be seen on a spectrum of downstream (closest to harm) to upstream (furthest from harm). More downstream audiences can be considered lower in prevalence but higher in severity, while upstream audiences are higher in prevalence but lower in severity. However, all five of these objectives ladder up to the overarching goal of counteracting the potential interest in terrorist and violent extremist groups and content.


1. **Deter** individuals from exploiting digital platforms to inflict harm and committing acts of VE.
2. **Intervene** with individuals who may pose a risk of VE and encourage them to disengage.
3. **Prevent** individuals from joining VE groups and address push/pull/personal factors.
4. **Build** primary resilience to VE radicalization.
5. **Empower** bystanders to prevent / counter VE.

## Strategies and Tactics to Achieve Positive Intervention Objectives

The section above sets out what positive interventions hope to achieve. The strategies categorize **how** they achieve this, and the tactics categorize concrete intervention types. Some strategies may help to achieve more than one of the objectives.

The tactics have been assembled based on the case studies brought forward by GIFCT members, theoretical interventions proposed in P/CVE literature, and interventions from neighboring harms that take a public health approach to prevention, such as child safety.

## Strategies

1. Provide direct support
2. Rebalance perceived cost / reward
3. Counter VE propositions
4. Build resilience to radicalization
5. Supply less extreme alternative
6. Raise awareness and skills

| Objectives / Strategies | Provide direct support | Rebalance perceived cost / reward | Counter VE propositions | Build resilience to radicalization | Supply less extreme alternative | Raise awareness / skills |
|---|---|---|---|---|---|---|
| Deter harm | | | | | | |
| Intervene | | | | | | |
| Prevent | | | | | | |
| Build resilience | | | | | | |
| Empower bystanders | | | | | | |

## Tactics

All concrete intervention tactics fall under these six strategic approaches:

### Provide Direct Support

Tactics that fall under this approach include de-escalation from violence which generally takes the form of either self-help content or one-to-one outreach on social media platforms. Some classic examples of programs that aim to provide direct or indirect support to those seeking to disengage from violence are the EXIT programs in Sweden, Germany, the United Kingdom, and the U.S. (Case Study 6).

Providing direct support can also take the form of bridging to support services for those at risk of radicalization and/or recruitment. These include digital communications to increase engagement with disengagement services, counseling or crisis support, and support for bystanders to intervene when friends and family are displaying radicalizing behaviors. For example, Facebook's partnership with EXIT U.S.A. aims to facilitate engagement with existing disengagement services (Case Study 11).

**Case Study: Counter Conversations**

Positive interventions can also provide opportunity with more direct engagement with individuals than the use of multimedia content can allow. The ISD Counter Conversations campaign involved the application of offline counter radicalization interventions methods to an online context. Delivered on Facebook and targeting both the extreme right and violent jihadist ideologies, the campaign sought to provide an opportunity for individuals showing clear signs of radicalization to interact and engage with someone who could support them. This involved the deployment of a wide range of practitioners to conduct the interventions, including former members of extremist movements and counsellors to directly message those who were ascertained to be at risk of radicalization. For those who were perceived to be further downstream than others, the practitioners would encourage their contacted individuals to engage with other services in order to further disengage from the radicalization process. That the intervention providers received a relatively high response rate showed that there is potential to scale this form of direct online engagement beyond Facebook and to other social media providers.

## Rebalance Perceived Cost/Reward

Tactics that fall under this approach include the flagging of violent and harmful content and raising awareness of the effects on victims. One form of this is to remind users of the social penalties of their actions such as Twitter's warning to users about sharing and reading misleading information (Case Study 13). Although it should be noted that while this system has been limited to adjacent fields such as misinformation and propaganda, it has not been tested in the context of extremism. Another form this tactic took is to remind the user of the illegality and potential consequences of their transgressions. Google's OneBox warning system in relation to child abuse (Case Study 12) is a good example of this in another field which reminds individuals that they are being observed and raising awareness of the penalty for such acts.

## Counter Violent Extremist Propositions

Tactics falling under this approach include counter narratives, trojan content, and fact checking. Counter narratives refer to content that directly undermines a violent extremist ideology, narrative, organization, leader, or violence itself. There are many examples of this tactic including the Mythos Lab and East India Comedy's anti-ISIS collaborations (Case Study 25). Trojan content refers to content that is ostensibly sympathetic or aligned to violent extremism, but that reduces trust in those organizations and ideas. Anonymous's UnManifest campaign called for online users to deface Anders Breivik's manifesto and create altered versions which mock him and his ideology (Case Study 20). Finally, fact checking involves an authoritative review of violent extremist statements with well-sourced and balanced refutations of arguments. Civic Fab's "What The Fake" is a good example of this tactic (Case Study 31).

## Build Resilience to Radicalization

This category consists of tactics aimed at audiences considered to be upstream or midstream. Inoculation for example aims to proactively raise audience mental defenses to counter future violent extremist messages. Another method is to address grievances, which encourages audiences to empathize and constructively suggest non-violent solutions.

**Case Study: Nigeria United**

It is clear that different audiences which are at different stages of the radicalization process require different alternate narratives. Zinc Network's "Nigeria United" campaign aimed to facilitate inter-faith cooperation through football for those vulnerable to Boko Haram. Using a mixture of online activity, radio promotion, and offline events, the project emphasized shared identities and positive alternatives to extremism. This was aimed at a more upstream or midstream audience. More specifically, the project made use of precise localized targeting which encompassed the specific contexts of ethnic division and interreligious divisions at play in Nigeria. This campaign used the target audience's interests (in this Case the African Cup of Nations football tournament) as a means of constructing a credible preventative tool which the target audience could engage with.

Another method to build resilience is to tackle individual push, pull, and personal factors. This includes supporting mental health conditions, building self-esteem, addressing identity crises, reducing impulsivity, and promoting belonging. Moonshot's offering of mental health and social-grievance messaging to users searching for violent Islamist extremist content online in Indonesia is a good example of this (Case Study 8b). Finally, campaigns deploying this tactic may also provide education in the form of improving digital literacy and improving the user perception of an outgroup.

## Supply Less Extreme Alternative

CVE campaigns can also aim to provide a less extreme alternative to individuals at risk of radicalization. This could take the form of alternative narratives which promote hopeful or factual content and other perspectives. Extremist Hashtags such as #StopIslam have been flooded by CVE practitioners (Case Study 19) to instead promote factual content like infographics, URLs, and memes to undermine the narrative initially pushed by right-wing extremist groups. These alternatives could also take the form of different pathways for at-risk individuals to disengage. Another method is to simply provide safer options for users in the form of less dangerous but similar content to what was originally searched or posted. Campaigns using these tactics may simply aim to detoxify conversations by encouraging respectful communications or calming public discourse following a terrorist attack. Bodyguard AI (Case Study 32) seeks to achieve this via the use of artificial intelligence while HateAid (Case Study 33) sought to re-empower victims of hate speech so that positive voices are not silenced by extremists.

## Raise Awareness and Skills

There are tactics that are aimed at raising awareness and skills of users. These can involve encouraging users to report hateful or violent content and their associated individuals to law enforcement, tech platforms, or prevention platforms. EXIT DE's 'Hass Hilft/Hate Helps' program is a prime example of this (Case Study 34):

**Case Study: Hass Hilft/Hate Helps**
EXIT DE's Hass Hilft campaign targeted the extreme far right in Germany. It provided users with a means to flag hateful and extremist comments online and "rewarded" them with donations to charity. This dynamic encouraged users to flag comments they believed were hateful or extremist, as they might believe that it was funny to subvert the intent of such comments through effectively channeling the initial commenter's hate into a positive outcome. Over the course of the campaign over €10,000 was raised which was used to fund further NGO disengagement services. This success led the campaign to receive a lot of good press and interest in multi-stakeholder partnerships.

Raising awareness can also involve providing prevention resources, raising awareness of the drivers of radicalization, and promoting skills on spotting the signs of violent extremism and how to act. Educate Against Hate provides parents and teachers practical advice on protecting children from extremism and radicalization (Case Study 43).

## Findings

From Case Study submissions and interviews with GIFCT members, we have identified common lessons learned from the development and delivery of positive interventions. Many members faced shared struggles and have constructive advice for other stakeholders. These findings broadly fall into five categories: audience, strategic focus, delivery mechanisms, impact measurement, and multi-stakeholder collaboration.

This section details those insights provides tangible examples from case studies and makes recommendations for GIFCT stakeholders.

# 1. Audience

It is clear that practitioners understand their target audience but struggle to reach them effectively. Many seek behavioral-based proxies for known vulnerabilities to radicalization. However, there are consistent frustrations with the success of these approaches. For example, ISD's "Reparlons Jihad," "Et toi, le Jihad?" (Case Study 30) aimed to reach a distinct audience of individuals who may be considering or are already partially convinced by ideas linked to radical Islamism through the use of satirical cartoons. While the campaigners understood that ISIS exploited the suffering of Muslims in Syria to recruit new members, they had difficulty knowing whether the audience they reached was their intended audience. It was difficult to discern between those who were genuinely at risk of extremism and those who were simply concerned by the suffering in Syria. Similarly, "Extreme Dialogue" (Case Study 3), a series of videos aimed at undermining extremist narratives through the use of short, emotive videos, targeted some downstream audiences across Facebook, Twitter, and YouTube. However, a CTR (click through rate) of 0.5% suggested that the campaign had trouble reaching and influencing its target audience.

Many practitioners report reaching an audience that is too wide for the objective, prompting low impact, backlash, or compromises on strategic focus. Practitioners who focus on training, research, resource development, and social innovation to counter extremism, such as HIVE Pakistan (Case Study 38), find it difficult to define a distinct audience, set achievable goals, and evaluate measurable, effective outcomes.

### Case Study: Search-Based Targeting

Paid search campaigns serve as an example of the potential of search-based targeting in CVE campaigns. Moonshot's Redirect Method campaigns used Google Ads technology to serve CVE content to users searching for violent jihadist and violent far-right content. This was used for a multiplicity of aims and contexts. These included offering mental health and social grievance messaging to users searching for violent jihadist content in Indonesia, serving CVE content to U.S. users searching for election and conspiracy theories, and mapping how Canadian users engage with violent far right & Islamic extremism online in English, French, and Arabic languages.

The utilization of advertising technology enabled these campaigns to be accurately targeted at their intended audiences. This utilization also enabled the operators of these campaigns to conduct thorough audience analysis and impact measurement. In the Case of the Redirect Method's deployment in Indonesia, this enabled the campaigners to conduct analyses pertaining to the effect of gender or age on the likelihood to engage, as well as conduct multiple comparable campaigns to more clearly understand the impact of said campaigns. Paid search campaigns allow for practitioner analysis to move past engagements and into watch time and click through rates.

Search-based targeting appears to have helped practitioners achieve greater success at reaching downstream audiences. Practitioners seeking to reach higher volumes of bystanders and upstream audiences have favored other approaches.

Practitioners occasionally conflate the ability to reach the target audience with the effectiveness of delivering impact. Regardless of targeting methods, inconsistencies around evaluation make it difficult to compare the impact of projects with different audiences.

## 2. Strategic Focus

This section outlines the strategic focuses of different positive interventions. There are roughly four different buckets of audiences related to radicalization, for which positive interventions are delivered, each of which requires distinct strategic focuses: upstream, midstream, downstream, and bystanders. The desired outcomes will look different across each of these stages and will require targeting different audiences and tailored delivery mechanisms.

### Upstream

Upstream approaches are generally preventative and aimed at broader audiences. These approaches can include building resilience to violent extremist or terrorist narratives or promoting social norms such as CST's "dontblameme" campaign, which focused on promoting the value of collective human identity over religious division (Case Study 35). Upstream approaches can also focus on raising awareness or addressing widely-held grievances. For example, EXIT U.K.'s creation of short video stories addressed how day-to-day deprivations and grievances can be compounded and preyed upon by radical right extremist organizations (Case Study 24).

### Midstream

Midstream approaches aim at a narrower grouping of individuals, typically those who are more at risk of radicalization but not yet recruited to a violent extremist organization or radicalized to violence. Methods within this approach include addressing vulnerabilities and building resilience such as in the Zinc Network's "Nigeria United" campaign (Case Study 39). Other methods include inoculating against violent extremist narratives, providing alternative narratives and pathways, and undermining confidence in the group, ideology, and its leaders.

### Downstream

Downstream campaign approaches aim to directly rebut, refute or counter narratives of violent extremist or terrorist groups. They also aim to counter the justification, incitement, or glorification of terrorist acts. Multiple downstream campaigns have encouraged engagement with disengagement services such as EXIT DE's disengagement program

(Case Study 6), and Moonshot's Redirect Method, which targeted those interested in armed militias following the January 6 Capitol Hill riot to engage in calming content and mindfulness practices (Case Study 8).

### Bystanders

Campaigns can also target the peer network or broader community that surrounds at-risk individuals. The aim is generally to encourage bystander interventions on an individual believed to be at risk of radicalization and/or recruitment.

**Case Study: The Stabilization Network's Women in Syria Program**
CVE campaigns can also take the form of creating online communities and spaces for vulnerable individuals and bystanders to interact safely within. The Stabilization Network's Women in Syria Program created such a community which aimed to solidify an online community that was resilient to extremism and challenged some of the gender stereotypes found in extremist messaging. This group was highly active with up to 1000+ of user generated media per month. Having traditionally been left out of discussions about how to combat extremism, women reported feeling more confident in intervening to protect their children from violent extremist manipulation. As the content is user generated, these kinds of online communities are sustainable once the initial campaign has run its course.

Practitioners are under-resourced and often aim to achieve more than is possible with a single positive intervention. This can lead to a reduction in strategic clarity and impact. Many areas in the theoretical framework of positive interventions are not covered by practitioners' case studies. Warning messaging that reminds those trying to access extreme content of its illegality and harmful social impact on the searcher has been effective in the areas of cybercrime and child abuse but has not been deployed in a violent extremism context. Additionally, inoculation messaging for violent extremism is a relatively new and unexplored area, although there is recent and increasing research on this.

Many insights from academic and tech sector research or experiments are not being used by practitioners in positive interventions. These include issues where evaluation is not the focus of the campaign or done robustly. Examples of this include the previously discussed EXIT DE's "Hate Helps" campaign as well as others featured in the Audience section. Additionally, many of these campaigns lack the technical and advertising skills necessary to reach an extremist audience.

## 3. Delivery Methods

This section outlines the delivery mechanisms used by practitioners to deliver positive interventions, the advantages and disadvantages of these approaches, and recommendations for consideration by GIFCT stakeholders in future intervention activity. The delivery mechanisms used by practitioners fall under one (or several) of the following categories: organic, paid advertising, private messaging, group messaging, and pop-outs/safety notices.

### Organic

Organic interventions have been delivered across a wide range of mechanisms (aside from "standard" organic posting on channels including disruptive flooding (attempts to confuse, exhaust, or inhibit connections in the absence of algorithmic solutions) and leveraging of search engine optimization (SEO) algorithms (optimizing organic search results to show positive alternative content). While many practitioners report significant reach and media coverage from these approaches (#StopIslam), media coverage and reach do not necessarily indicate impact. These types of interventions are hard to measure and there is evidence that extremist networks are more tightly integrated and therefore more resilient to waves of hashtag activism. There is also evidence of diminishing returns with this approach – there was less counter speech against Islamophobia and minimized reach/impact after each successive Islamist extremist terrorist incident in the U.K. in 2017.

When designing interventions relying on organic reach as the primary delivery mechanism, practitioners should consider both channel selection and viral hooks. Channel selection refers to selecting a channel that has a high potential for organic reach is particularly important. Most interventions that have delivered high reach have used Twitter or leveraged SEO algorithms (or a combination of both) to achieve reach and scale of messaging. Engaging "viral" hooks refers to utilizing an angle (topical or not) that is likely to drive shares, audience participation, and subsequent media coverage. The Fighting Terror with Comedy campaign (Case Study 25) and the Brainwash and HI-SIS campaign (Case Study 26) did this through leveraging comedy as a viral hook and mainstreaming appeal, generating millions of views, dozens of press articles, and thousands of positive comments.

While strategic communications agencies have the expertise to deliver interventions using the full suite of digital marketing tools (including paid advertising), practitioners that lack this expertise struggle to achieve reach/impact – particularly if they are reliant solely on organic reach. Practitioners should be encouraged to use the full range of resources at their disposal.

## Paid Advertising

Paid advertising provides one of the best routes to ensuring positive interventions reach the correct audience as well as enabling effective measurement of performance. Positive interventions have been delivered using a wide range of paid digital advertising, including paid search (targeting and intervening in searches for extremist content), Pre-roll advertising (placing ads against specific content or based on wider targeting parameters), paid social (advertising on social media channels), and display advertising.

The impact of practitioners' use of paid advertising is limited by budget. While some ad grants/ credits are available to CSOs, strategic communications agencies have to rely on client budgets to deliver paid interventions, which can be very limited (e.g. Zinc Network's Guide when using Google search – Case Study 7).

Self-serve digital advertising, such as on Facebook's Ads Manager or through Google Ads, does not include VE-related terms and groups as "interests" due to tech platforms' efforts to safeguard users and prevent extremist groups from abusing ads. While this prevents terrorist organizations from abusing social media / advertising platforms, it also restricts P/CVE practitioners from being able to reach target audiences with positive interventions. Practitioners using paid advertising are now reliant on broader targeting proxies that are less precise. In Search for Common Ground's GIRLS program (Case Study 28), practitioners collaborated with popular influencers to help promote messages of tolerance and pluralism among young women vulnerable to extremism. Without an accurate way to target these women, it was difficult for practitioners to be sure that these messages reached their intended audiences.

The same is true where search advertising has been deployed, as its effectiveness can be undermined by automated enforcement of platform policies. Programs like Moonshot's Redirect Method, ISD's Bing pilot, and Zinc Network's Guide Methodology (Case Studies 7-9) all use targeted search advertising to reach a specific audience searching for extremism online. These efforts have shown promise in engaging a range of at-risk individuals with alternative and supportive content. However, they can be disrupted or disabled completely due to platforms' enforcement practices against their terms of service.

## Private Messaging

Private/direct messaging of individuals expressing support for violent extremism appears to be a highly effective means of delivering downstream positive interventions. ISD's Counter Conversations program (Case Study 29) provided promising evidence that a solid proportion of individuals expressing support for violent extremism online can be identified at speed and scale and encouraged to engage with online intervention providers on a sustained basis. However, this is a resource-intensive approach that requires investment and further development in order to scale.

Private messaging interventions have shown promising results, but they have not been deployed in all contexts, have not been scaled up, and many variables have not been proven.

## Group/One-to-Many Messaging

Group and One-to-Many messaging generally fall into two categories: the use of credible voices such as influencers or formers and the use of events. There are a number of recommendations that can be pulled from each of these.

The choice (and credibility) of messenger is highly important in delivering effective positive interventions. Practitioners have experimented considerably with the most effective messengers for positive interventions, and this has usefully shaped future investment and practice. Campaigns that have used formers (EXIT formers – Case Study 6), Jamal al-Khatib formers (Case Study 22), Open Your Eyes to Hate (Case Study 41), influencers (e.g. footballers, imams in the Zinc Network football campaign – Case Study 42), other localized credible messengers (local hip-hop artists to create and deliver counter narratives debunking Islamist extremist propaganda), and the engagement of hyperlocal influencers and peer-to-peer networks (Case Study 37) have achieved impact.

While the credibility of the messenger is important, campaigns funded by governments run the risk of being undermined if they are not attributed and then subsequently "exposed." Campaigns such as Ummahsonic (Case Study 40) needed to seem separate from the government, but at the same time needed attribution to the donor or practitioner to avoid risks.

Some of the most impactful initiatives discussed here linked digital communications with offline service provision. Events provide a highly effective way of delivering meaningful positive interventions to upstream audiences in a credible, community-led setting. Zinc Network's use of football training (in both Nigeria and the U.K.) as a vehicle to build resilience, self-esteem, and promote pluralism delivered significant impact – the U.K. project saw message retention scores of over 90% and increases of up to 20% in feelings of self-esteem and belonging (Case Study 39). And ISD's Counter Conversations program showed that this is relevant for disengagement too (Case Study 32). Nafees Hamid's work suggests that small-group dynamics (rather than online mass communications) are more conducive to the spread of extremism. Events, small groups, and communities are therefore key in delivering positive interventions.

Pop-out/warning messaging on tech platforms (e.g. self-harm support messages, child sexual abuse warnings, and misinformation warnings) exist, although there has been limited independent evaluation of their efficacy.

## 4. Impact Measurement

Measuring the impact of positive interventions remains a crux for practitioners delivering this work, in large part due to the need to balance data access and data privacy, and to link online engagement with long-term offline behavioral outcomes. This section sets out two key challenges.

**Practitioners use digital metrics to evaluate campaigns but struggle to find meaningful impact data**

Many counter-narrative campaigns rely solely on basic metrics to approximate efficacy, instead of leveraging the full suite of available metrics. Counter campaigns like Brainwash and HI-SIS's comedic videos and Exit U.K.'s video stories (Case Studies 24 and 26) generate views, but other available data points like average view duration or percentage of video watched are not included. As such, basic metrics are often used as a shorthand for success, rather than interrogated to shed light on the extent of a campaign's effectiveness.

Longer-term analysis is often neglected due to the difficulty in capturing this digitally. The analysis of the #StopIslam hashtag flooding is an example of where the impact was better judged as a result of measuring longer-term effects. They used a nodal analysis of the information-sharing networks around #StopIslam and found that the original far-right movement using the hashtag was better at outlasting individual waves of activism generated from recently-mobilized networks. They highlight that there was also little counter speech re-mobilization in the face of online attacks on Islam following the 2017 terror attacks in Manchester and London. By reviewing the longer-term effects alongside digital metrics, these additional analyses highlight the transient elements of the #StopIslam counter campaign, and in turn allows the impact to be more effectively measured.

The most promising campaigns – like Facebook's partnership with Exit U.S.A. – aim to combine digital metrics with additional data on service engagement and recidivism to better approximate impact. More effective measurement of impact occurred when projects had a baseline against which to compare their results. Moonshot's testing of mental health and social-grievance messaging to users searching for violent Islamist extremist content online in Indonesia found that engagement with messages addressing loneliness was 128% greater than with traditional counter-narrative messaging.

**Practitioners struggle to control the controllables in positive interventions, rendering impact data less meaningful**

As positive intervention campaigns occur "in the wild" rather than in controlled lab conditions, practitioners find it difficult to have a control grouping to compare their results against. This reduces the meaningfulness of the collected impact data post-campaign. The work from the International Centre for the Study of Violent Extremism (ICSVE) is a great example of the insights possible when the variables are controlled (Case Study 5). Their campaign used videos of formers to provide counter narratives to ISIS ideology, narrative, leaders, and strategies. Effective testing generated key insights on the optimal length of video and most effective messenger.

## Recommendations

GIFCT's foundational goals include "enabl[ing] multi-stakeholder engagement around terrorist and violent extremist misuse of the Internet and encourag[ing] stakeholders to meet key commitments consistent with the GIFCT mission;" and "promot[ing] civil dialogue online and empower[ing] efforts to direct positive alternatives to the messages of terrorists and violent extremists."

Drawing in part on the findings above, the following recommendations are all offered with a multi-stakeholder approach in mind.

## Information Sharing

- There should be greater information sharing around best practices for both conducting and evaluating positive interventions. When positive interventions are shown either to succeed or to fail, information about the design and implementation of those interventions should be shared as widely as possible, including how best to reach different audiences and communities. Although there are legitimate reasons for withholding information, as this report has illustrated it is possible to share lessons learned while also respecting important privacy and intellectual property considerations.
- Similarly, there should also be greater information sharing about how best to evaluate positive interventions. Practitioners should work with academics and the research community to develop standard quantitative and qualitative methodologies for the evaluation of positive interventions. The methodologies should balance inferential rigor with ease of use and implementation so that firms and practitioners can carry them out without necessarily having methodological expertise themselves.

## Support and Capacity Building

- Governments, tech, and civil society should consider opportunities to better support practitioners. Depending on the sector's expertise, this could include practical support such as direct or third party capacity building, making research available to practitioners, or the provision of funding to support campaigns. In doing so, governments and tech should increase awareness of already available resources.

- GIFCT, governments, and tech should consider enhanced ways to support research into the efficacy of positive interventions. This could include the development of an evaluation framework, controlled evaluation of different positive interventions, and identification of different audiences and communities.
- All sectors should consider further opportunities to support and enable the delivery of positive interventions across online and offline environments, with a particular focus on audience definition, effective delivery, and measuring impact.

## Linking Online and Offline Environments

- Where users may be engaged in positive interventions in online spaces, all sectors should consider ways to support and facilitate users that may be at risk of radicalization to positive interventions and support that is delivered offline. There are opportunities to explore further best practices in this area.
- Consider ways to facilitate effective hand-offs between the online and offline spaces, particularly in cases where there is an escalation of risk. Where risk escalates, all sectors should share information effectively and government agencies are best placed to take over responsibility for potential risk of radicalization and/or threat disruption.
- Governments should develop agreed-upon definitions of commonly used terms in areas of terrorism and violent extremism to assist other sectors to navigate the terrorist and violent extremist environment with consistency and shared understanding.
- Focus efforts on prevention space and recommendations for users that may otherwise unwittingly be led to radicalization.

## Investing in Innovation

- Many of the positive intervention strategies and tactics identified in Part I remain under-utilized. More consideration should be given to how best to fund and implement programs across the full range of positive intervention strategies available.
- The current knowledge base for positive interventions is based on programs that were conceived years ago before the emergence of new platforms and features. In line with the Christchurch Call workstream on positive interventions, greater consideration and support should be given to new and innovative strategies that

new and emerging technologies have made possible.

## Conclusion

The GIFCT CAPPI WG members recognize the distinct strengths and capabilities of civil society, academia, government, and the technology industry. Leveraging those strengths will substantially increase the likelihood of successful interventions which can be best achieved by taking a multi-stakeholder approach. The trust that was generated among the multisector participants of the CAPPI WG offered a new opportunity to explore lessons learned and best practices across sectors. This is something the GIFCT should look to build upon when considering how to incorporate the recommendations of this WG.

## Acknowledgments

The multi-stakeholder participation in this WG has yielded significant benefits. GIFCT would like to thank the CAPPI participants from all sectors for contributing case studies and providing such a comprehensive mapping that led to these recommendations. We also thank participants for permitting the publication of case studies in the annex, many of which are being shared publicly for the first time.

## Full list of participating individuals and organizations

| | |
|---|---|
| Aqaba Process | Microsoft |
| Brookings Institution, Chris Meserole (Co-Facilitator) | Mnemonic |
| Etidal | Moonshot |
| Facebook | Netsafe (NZ) |
| Global Partners Digital | New Zealand Government, Department of the Prime Minister and Cabinet (Co-Facilitator) |
| Google (Co-Facilitator) | Swansea University |
| Hope Not Hate | The Government of Ireland (Department of Justice) |
| Human Cognition | Twitter |
| Institute for Strategic Dialogue | United States Government (State Department and Department of Homeland Security) |
| Jihadoscope | Wahid Foundation |
| M&C Saatchi | Zinc Network |
| Maarif Institute | |

# Appendix

## Case Studies of Positive Interventions Provided by GIFCT Contributors

| No. | Case Study – About | Strategy | Delivery Vehicle | Lessons Learned |
|---|---|---|---|---|
| 1 | One-to-one pilot & program – attempt to bring disengagement best practices into the social media domain through messenger and to disseminate counter speech through direct messenger. | Provide Direct Support – one-to-one outreach | Private messaging function | Initial results show potential positive impact indicated by 10% sustained conversations. |
| 2 | Average Mohamed – cartoon series for Muslim Americans that aims to build resilience to radicalization by normalizing the struggles the TA may face, and promoting alternative narratives on identity and grievance. | Supply Less Extreme Alternative – Offer safer alternatives to at-risk users (narratives, pathways)<br><br>Raise Awareness – (reach family and friends to safeguard) | Organic reach<br><br>Facebook Ads | The extent to which a message countering Islamic extremist ideology was received differed across platforms: reach compared to engagement was highest on Twitter (1:14) compared with YouTube (1:304) and Facebook (1:141).<br><br>Sentiment suggests message was well-received: 305 comments; 66% supportive; the rest were "negative" or "unrelated."<br><br>Longer-term impact beyond engagement metrics was difficult to ascertain and was not a focus of the campaign.<br><br>Experienced spam messages when their videos went viral. |
| 3 | Extreme Dialogue – guided navigation for young people (and those who work with them) to understand radicalization and prevent it in themselves and among their friends. | Raise awareness / skills – Offer safer alternatives to at-risk users | Organic reach | Evidence of engagement with counter extremist narratives – 65% of videos watched on average.<br><br>Analysis of audience demographics suggests that YouTube generated a more age-appropriate audience.<br><br>Low clickthrough rate (CTR) suggested that the ads might have either reached the wrong audience or relied on unappealing advertisements. |
| 4 | Harakat-ut-Taleem Campaign – counter Taliban recruitment narratives in Pakistan. Aimed to discourage young Pakistanis in the U.K. and Pakistan from joining an Islamist extremist group by highlighting the negative consequences. | Rebalance perceived cost / reward<br><br>Counter VE Propositions | Organic reach | Harakat-ut-Taleem videos targeted in Pakistan (without English subtitles) received higher viewer retention rates compared to the English subtitled versions targeted in the U.K..<br><br>For Harakat-ut-Taleem we ran advertisements for "engagements" to test out these different options. There appeared to be a clear indication that video view ads are better at fueling engagement than engagement ads.<br><br>Building a brand and following from zero was difficult and lacked credibility (Harakat-ut-Taleem went from no presence on social media to 116 Facebook likes, 6 YouTube subscribers, and 62 Twitter followers.)<br><br>Harakat-ut-Taleem's videos performed better in Urdu in Pakistan than they did with English subtitles in the U.K., demonstrating the ability for counter narratives coordinated in the U.K. to have a global impact.<br><br>Targeting relied on keywords / interests which are no longer available. |

| | | | | |
|---|---|---|---|---|
| 5 | ICSVE – interviews with terrorists and extremists, edited into social videos to provide counter narratives to ISIS ideology, narrative, leaders, strategies. | Counter VE propositions | Organic reach | Several case studies in which prisoners have changed their opinions based on these interviews.<br><br>ICSVE has experimented with lots of variables within their content: on FB One-minute videos yield higher viewership than longer videos;<br>For M&F audiences, engagement is greater with white, female messengers (sympathetic messenger), and with uncovered face, name used (trustworthy messenger). |
| 6 | Exit Groups (Exit Sweden, Exit Germany, Exit U.S.A., Exit Australia, etc.) – disengagement organizations that aim to support far-right extremists to leave the movement. | Provide direct support – De-escalation from violence, bridging to support services<br><br>Alternative narrative (showcasing formers, providing hope for redemption) | Organic reach | Assessments of longstanding Exit programs like Exit Germany and Exit Sweden indicate success in exiting individuals from extremist movements.<br><br>Exit Germany: Since the year 2000 over 500 individual cases have been successfully finished with a recidivism rate of approximately 3%.<br>Newer programs like Exit U.S.A. indicate promise. In 2019, Exit U.S.A. managed over 200 Cases. Roughly half were seeking services for themselves and half were seeking services for a loved one. Life After Hate (which runs Exit U.S.A.) has partnered with Facebook on its Redirection initiative, and will launch the Redirect Method in 2021 to redirect users to their website. |
| 7 | The Guide Methodology – targeting methodology, leveraging SEO on low-prevalence, high-risk terms to direct users toward pre-existing counter-radicalization resources. | Digital marketing approach that could be used for all strategies, content curated from elsewhere | Targeted advertising | Medium-intent audiences are open to interventions and being engaged with by alternative narratives.<br><br>Searches for banned publications suggest audiences at an early stage of the radicalization journey.<br><br>The most effective campaigns using Guide point to multiple landing pages, with content on each page aligned with (and relevant to) the search terms driving traffic to those pages.<br><br>Islamist audiences responded the most to ad copy that mimicked Islamist communications . Far-right audiences responded the most to ad copy that addressed an underlying grievance or pervasive local narrative.<br><br>Lack of relevant content to point users to.<br><br>Lack of client appetite to commission relevant content.<br><br>Need for websites to host content and align with search advertising user journey. |

| | | | | |
|---|---|---|---|---|
| 8a | The Redirect Method – targeting methodology, leveraging SEO on low-prevalence, high-risk terms to direct users toward pre-existing counter-radicalization resources. | Digital marketing approach that could be used for all strategies, content curated from elsewhere | Targeted advertising | External evaluation demonstrates efficacy of implementers to use advertisements to effectively expose individuals to content that offers alternative narratives.<br><br>In an assessment of deployments in over 25 countries, deployments in North America and Europe see users click on these ads at a rate on par with industry standards. Asia and the Pacific have seen the highest CTRs.<br><br>Users seeking information specifically on extremist groups (for example, seeking information about joining a group) tend to be disproportionately likely to engage with alternative content offered by the Redirect Method.<br><br>Searches for extremist content spike in the immediate aftermath of terrorist attacks. Engagement with Redirect campaign content also spikes in the aftermath of these attacks, indicating the appetite for redirection during crises.<br><br>Initial published results show higher engagement with social service content among those at risk of violent extremism.<br><br>Robust "allow listing" process required by tech companies to manage access by trusted 3rd parties to key search terms.<br><br>Process required to manage risks associated with tech companies profiting from advertising to violent audiences.<br><br>Suspensions: As a result of no streamlined whitelisting for campaign providers, campaigns can be prone to regular suspensions and require intensive maintenance.<br><br>Limited ability to evaluate long-term impact given lack of user-specific data. |
| 8b | The Redirect Method – offering mental health and social-grievance messaging to users searching for violent Islamist extremist content online in Indonesia. | Build resilience to radicalization | Targeted advertising | Compared to ads offering counter-narrative content, Indonesian audiences searching for violent Islamist extremist content online were 16% more likely to engage with employment support.<br><br>Compared to ads offering counter-narrative content, audiences looking to join or engage with violent groups were 128% more likely to engage with loneliness campaign and 41% more likely to engage with employment campaign.<br><br>Male users were more likely to engage with ads offering support for depression and stress and were least likely to engage with the ideological counter-content.<br><br>Conversely, users identifying as female were most likely to seek out support for anger, loneliness, and unemployment.<br><br>Users over the age of 55 were most likely to engage with support for depression compared to their younger counterparts, who were far more likely to engage with employment support. |
| 8c | The Redirect Method – U.S. election & conspiracy theories. | Build resilience to radicalization<br><br>Supply less extreme alternative | Targeted advertising | 1,330 engagements with a campaign designed to draw people away from violence.<br><br>33 hours watched for videos promoting calm & mindfulness techniques.<br><br>6% CTR on ads related to moving on from anger when shown to those searching for Q-Anon.<br><br>Important to consider deep vetting processes for content creators. |

| | | | | |
|---|---|---|---|---|
| 8d | Canada Redirect – mapping how Canadians engage with violent far-right & Islamic extremism online (English, French & Arabic languages). | Build resilience to radicalization<br><br>Supply less extreme alternative | Targeted advertising | 2,583 clicks and 3,960 video views, suggesting multiple views for every click. Content was viewed for over 58 hours in total.<br><br>Users seeking information on VFR extremist groups are disproportionately likely to engage with Redirect content.<br><br>Music provides a unique opportunity to keep the attention of VFR at-risk users given the higher average watch time.<br><br>Searches & ads in Arabic, although making up a small portion of overall searches, produced significantly better CTRs (8%), suggesting Arabic as a relevant option for IE audiences.<br><br>Risk assessment and risk-rating are critical to conducting ethical strategic communications and their rigor can be evaluated. Continued refinement of risk indicators key to success (i.e. should not be a static project).<br><br>Customization and targeting do not guarantee success. This is not digital marketing – the desired outcome is far more complex and so what works in theory for digital marketing may not work in all CVE contexts. |
| 9 | Bing's Ads Pilot – serve compelling counter narratives when people search for terrorism and extremism related content on Bing. | Counter VE propositions | Targeted advertising | Make use of all the metrics that are available via search-ad platforms (e.g. keyword impressions can be used to monitor the "frequency" of specific searches over time). This can then be used to determine whether a keyword approach needs to be adapted.<br><br>Also trial different types of content – link ads to videos and landing pages. Videos with a commenting function are useful to gain a sense of how users are engaging/receiving the content. With sites like YouTube, how many users reached a video via placed ads on Search Engines can also be determined.<br><br>Language of ads should be succinct and appeal to the narrative they are trying to counter. Do not dismiss the narrative in the ad as that will not draw users in. Prompt users with a call to action like "come learn more here" or something similar.<br><br>Monitoring impact beyond just reach and surface-level engagement (e.g. click-throughs) is difficult in this methodology. This can be avoided by uploading content where engagement can be monitored (e.g. where users can comment on the content, or where they are prompted to dial a number). Some form of online to offline framework is also a useful way to monitor actual individual-level impact, but this is difficult to arrange and carries lots of risks with it. |
| 10 | YouTube's "Featured video" Redirect – when people search for certain keywords on YouTube, it displays a playlist of videos debunking violent extremist recruiting narratives. | Counter VE propositions – content curated from elsewhere | Promotion of "Featured" content | Requires active updating of redirection content to ensure safe and relevant to at-risk audiences.<br><br>Requires regular maintenance of a keyword database. |

| | | | | |
|---|---|---|---|---|
| 11 | Facebook's signposting initiatives – These interventions connect people who search for terms associated with white supremacy on Facebook Search to resources focused on helping people leave behind hate groups. | Provide Direct Support – signposting services to at-risk users | Warning message | Independent evaluation found the pilot was broadly successful. This conclusion is based primarily on the fact that, during the three-month pilot, 25 Facebook users who initially sought to engage with violent extremism on the platform instead received some form of support from one of the delivery partners.<br><br>This demonstrates that the program is able to successfully link high-risk individuals with support, and help turn a passive search for violent extremist content into an active conversation. Upstream metrics complement this finding.<br><br>Tens of thousands of Facebook users were offered help by the safety module, of which thousands accepted and went on to engage with supportive content on delivery partner sites. Of these, 25 individuals chose to begin a conversation.<br><br>Facebook can use its search module to reach "low-prevalence, high-risk" audiences.<br><br>The initiative successfully created friction between the search for a white supremacist and/or neo-Nazi community and a positive result, and to some extent functioned as the conduit between high-risk individuals and their respective delivery partners, which has helped to turn some passive searches into active conversations. |
| 12 | Google's OneBox safety warnings – a child safety warning message | Provide Direct Support – signposting services to at-risk users | Warning message | https://www.pcworld.com/article/2064520/google-to-warn-users-of-13000-search-terms-associated-with-child-pornography.html |
| 13 | Twitter warning to slow spread of election misinformation | Rebalance perceived cost / reward: Deliver Warning | | https://www.npr.org/2020/10/09/922028482/twitter-expands-warning-labels-to-slow-spread-of-election-misinformation |
| 14 | Ahul-Taqwa magazine – Inspire/ Dabiq-like counter-content magazine, publicized and released on Telegram. | Counter VE propositions – trojan content | One-to-Many group messaging, on Telegram | Advocates suggest apparent credibility of messenger is effective, but little evaluation is evident.<br><br>Difficult to judge efficacy, specifically in regards to audience reaction to trojan delivery.<br><br>Promotion & amplification of counter-content runs the risk of being removed. |
| 15 | WAVE 100 in Nigeria – Digital PVE project mobilizing women to build community resilience via WhatsApp. | Build resilience to radicalization<br><br>Raise awareness / skills | Organic reach; targeted advertising; private messaging | Piecemeal evaluation, but some significant numbers reached & engaged.<br><br>Long-term change difficult to quantify. |

| | | | | |
|---|---|---|---|---|
| 16 | Extreme Lives – Running since 2017, the project explores the personal insights and stories of people who have been affected by violent extremism and of those exploring issues of identity, ethnicity, diversity and inclusion, focused in Asia. | Supply less extreme alternative<br><br>Raise awareness / skills | Organic reach on Facebook; media interviews; Facebook ads | |
| 17 | Resiliency Initiative – a partnership between Facebook and The Asia Foundation which aims to promote tolerance, strengthen interfaith and inter-ethnic understanding, and counter violent extremism by helping to build resilient communities across APAC. | Supply less extreme alternative<br><br>Raise awareness / skills – Positive alternatives, addressing grievances and building skills to tackle intolerance and discrimination | Capacity building of digital skills – then YouTube, Facebook campaigns (organic and paid) | |
| 18 | The Stabilization Network's female-specific P/CVE comms in the Levant – creation of a digital community, with user-generated content and online learning modules. | Provide Direct Support – conversations<br><br>Counter VE Proposition<br><br>Build Resilience to radicalization – inoculation and education | Private messaging | The impact of the project was the creation of a highly active digital community, with up to 1000+ user-generated pieces of content per month. Dialogue was successfully generated and curated among vulnerable women, and sustainably functions one year after project closure and is completely user-led with no external material support.<br><br>A closed female group among the target audience also completed online learning modules designed to promote agency and build individual resilience.<br><br>Female audience members also reported feeling more confident about engaging in decision-making in the home and civic life, reporting anecdotal incidents of change, and more confident in their ability to protect their children from violent extremist influences, particularly with regard to manipulation, brainwashing, and trauma.<br><br>The results are also in line with one of the good practices on women and CVE recommended by the GCT, namely that women "are well placed to effect change, especially at the community-level, and should be empowered to take ownership in the development and delivery of inclusive narratives to violent extremism and terrorism."<br><br>Research foundational to informing communications: literature review to confirm gap in understanding; sentiment analysis; network analysis of female Arabic Twitter accounts; interviews with female activists.<br><br>Lack of female-specific research in regards to P/CVE campaigns – there is a need for increased research into the use by vulnerable females of digital media, expanding to include dark social (particularly WhatsApp and Telegram), but also Facebook Messenger (bots). Learnings should be tested in other comparable areas where vulnerable audiences are located, namely Al Hol and similar locations. |

| | | | | |
|---|---|---|---|---|
| 19 | #StopIslam Twitter Campaign and Counter Campaign (2016) – the hijacking of an extremist hashtag by counterspeech activists. | Counter VE proposition – Counter narratives, disruptive flooding, Alternate narratives, Detoxify conversation, Calm public discourse | Organic reach, Hashtag hijacking | The top-shared tweet of the 2016 #StopIslam Twitter campaign and counter campaign came from the counter narrative side, with a top-shared tweet defending Islam being retweeted 6,643 times versus the top tweet attacking Islam being shared 1,500 times.<br><br>Geographically, counter narratives were shared mostly by tweeters in the U.K. (74% of the U.K. sample) and the MENA regions (86%)<br><br>Campaigns generated further free publicity. 22 of the top 100 Twitter accounts disseminating information about the hashtag were notable and more established media organizations. 64% of these institutions reported on the counter-narrative. According to Poole et al. (2019), this adds further weight to the argument that the hashtag was successfully appropriated by a counter-movement in order to "gain visibility for anti-racist, inclusionary discourse."<br><br>Effects of such counter narratives do not last – A nodal analysis of information-sharing networks, for example, found that right-wing extremist activist networks were more tightly integrated and better established when compared to the #StopIslam counter community – meaning that such movements were better at outlasting individual waves of activism using the #StopIslam hashtag. For example, there was little counter speech re-mobilization in the face of online attacks on Islam following the 2017 terror attacks in Manchester and London. This leaves space for better coordination at the activist level and more organized attempts by NGOs and CT officials. |
| 20 | Anonymous' UnManifest (July 2011) Campaign – a cooperative effort from the online hacktivists, Anonymous digital disruption to bury Breivik's "manifesto" at the bottom of search engine results. The campaign called for online users to deface Breivik's manifesto and create altered versions that use humor to mock its author and discredit his violent ideology. | Counter VE proposition – through counter narratives, calming public discourse, using humor, and detoxifying the conversation | Organic reach, Hashtag hijacking | Use of humor is a novel way to counter extremism. It was also an inversion of the process of "versioning" after the Christchurch attacks five years later, where radical right sympathizers and those bent on propagating the terrorist's manifesto created duplicated and different copies to avoid removal.<br><br>The campaign created quite a media stir with the story of the campaign being carried on several online news sites.<br><br>Not known how many at-risk individuals were affected by this.<br><br>Not known whether there was a 'backfire effect' – whether this encouraged more far-right activity in response. |

| | | | | |
|---|---|---|---|---|
| 22 | Jamal al-Khatib – alternative narrative campaign through online street work. | Provide Direct Support<br><br>Counter VE proposition<br><br>Build Resilience to radicalization by reducing vulnerabilities, empowering Network, enhancing critical thinking and inner autonomy of target group through discussions, positively affect self-efficacy through call to action | Private messaging, One-to-Many messaging, Organic reach, Search-based targeted ads, Side-bar targeted ads, Newsfeed targeted ads, by appropriating Islamist extremist terrorist audio-visual codes | Successfully reached adolescents who are at risk of getting into contact with extremist content via search requests and optimized keywords.<br><br>Successfully established contact and initiated discussions of the issues raised in the videos.<br><br>Basing the content of our videos on material developed in narrative biography work proved a highly effective strategy as it helped the target audience identify with him.<br><br>Collaboration with young people and "formers" improved the quality and authenticity of the content. Both groups expressed their strong motivation to take action against extremist online propaganda and to participate in public debates about it, while also expressing their frustration that they rarely found opportunities to do so.<br><br>Our appropriation of audio-visual codes employed in material produced by extremist organizations such as so-called IS was also a significant factor in ensuring that our material caught and sustained the attention of our target audience. For videos to engage adolescents who are at risk of being influenced by the narratives of extremists, the right images had to be selected, the right music had to play in the background, and the narrator had to speak the right kind of language.<br><br>We are continually faced with the issue that the social media platforms (YouTube, Facebook, and Instagram) complicate the possibility to use paid ads.<br><br>During the first campaign, our Facebook account was blocked for a week, and in all three seasons, the option to advertise our videos was restricted on all platforms.<br><br>We experienced two coordinated efforts by extremists to spam our channels with propaganda disguised as harmless arguments. Handling these required a lot of resources, but we managed to handle them well.<br><br>Decentralized online setting and the limited time period available for building relationships imposed by the campaign structure make it hard to be certain how many viewers and commentators were inspired to engage in processes of self-reflection as a result of our work. |
| 23 | Britain Furst – grassroots parody of Britain First (2014). | Counter VE proposition<br><br>Build Resilience to radicalization | One-to-Many messaging, Organic reach | "Halal Sunglasses," "Muslamic Timepieces" and likening Britain First's uniform to bin liners each received 39 retweets/14 likes/4 comments, 21 retweets/19 likes/2 comments and 65 retweets/24 likes/5 comments respectively.<br><br>Britain First itself took notice and reported the group to Facebook.<br><br>It was the subject of a number of national news articles, inspiring successor accounts and initiatives by other online activists, even beyond the main cycle of protest by the group itself.<br><br>Difficult to ascertain how many at-risk audience members a) saw this, and b) were encouraged to move away from Britain First as a result.<br><br>Unknown whether this hardened the resolve of some members. |

| | | | | |
|---|---|---|---|---|
| 24 | Exit U.K.'s video narratives – personal testimony videos from individuals vulnerable to radicalization and their family members, encouraging disengagement. | Counter VE proposition<br><br>Build Resilience to radicalization<br><br>Provide Direct Support | Organic reach, presumably some kind of advertising, though the type is unspecified | The 3 personal testimony videos received many views in total – 116, 422, 17,080, and 4,496 views respectively.<br><br>How to get this content in front of the at-risk audience is a likely challenge. |
| 25 | Fighting Terror with Comedy in India (Mythos Labs x East India Comedy) – Mythos Labs partnered with East India Comedy, one of South Asia's most popular online comedy groups, to create a comedic counter-narrative video titled "I Want to Quit ISIS." | Counter VE proposition | Organic reach, Promotion via online media outlets | The video was hosted on the YouTube and Facebook accounts of the comedians, generating over 1.5 million views, dozens of press articles, and thousands of positive comments.<br><br>Difficulty in ensuring this reaches the at-risk audience themselves, or those already in extremist movements.<br><br>As such, difficult to judge success even with existing metrics that have been collected.<br><br>Press articles do not necessarily equate to success. |
| 26 | Brainwash and HI-SIS: Comedy & CVE in India & Indonesia – Partnership of movie stars and comedians in India and Indonesia to create comedy videos that counter gender stereotypes used by terrorists. | Counter VE proposition | Organic reach, delivery vehicle unconfirmed | The videos garnered over half a million views and unanimously positive press reviews, highlighting the importance of gendered counter narratives.<br><br>The audience needs better definition to make this a more effective campaign.<br><br>Press articles do not necessarily equate to success. |
| 27 | P/CVE in Indonesia via website – An online platform set up to create and promote alternative narratives for young Indonesians confronted with violent extremism. The idea was to train ten ustadz and ten former combatants in media communications and encourage them to become "credible voices" who could unveil the realities of violent extremism. | Counter VE proposition<br><br>Provide Direct Support | Private messaging | Led to engagement with Indonesians currently trapped in refugee camps in Northern Syria.<br><br>Female accounts tend to more successfully engage users.<br><br>Project highlighted the need to strengthen credible voices (both online and offline).<br><br>Platform's message targeting needed evaluating through feedback analysis.<br><br>Extra grassroots efforts also needed to encourage people to accept returnees/ deportees/former prisoners as they attempt to re-join communities.<br><br>The ultimate aim is to link people with suitable mentors offline, but they have so far encountered difficulty building the requisite trust.<br><br>Additional problems include their undercover accounts facing takedown measures from Facebook content moderators.<br><br>The savviness of users who seem adept at identifying suspicious accounts.<br><br>The potential security risk of meeting people offline before revealing their true intentions.<br><br>Female messengers can be problematic if a perceived romance develops. |

| | | | | |
|---|---|---|---|---|
| 28 | Search for Common Ground's 'GIRLS' (an acronym for Generating Indonesian Resilience and Leadership Skills) program, Indonesia – influencers such as food and travel bloggers are encouraged to incorporate messages of tolerance and pluralism into their content. | Build resilience to radicalization | One-to-Many messaging, Organic reach | Little evaluation present – would only work with influencers with 1000+ following though.<br><br>Audience could be better defined beyond women online.<br><br>Collaborating directly with young people is the only way to do this work. Government tends to be too slow and clunky to operate effectively in this space, which involves swiftly changing trends and specific language. Credibility is everything, and imitation is transparent. |
| 29 | Counter Conversations – Counter Conversations was a direct, online interventions project that used publicly available information to identify and engage users expressing extremist sentiment online. A team of trained intervention providers initiated outreach and engaged the users with constructive, personalized, and positive conversations to deter them from joining violent extremist groups. Intervention providers ranged from former extremists to survivors of extremist violence and counselors. | Provide Direct Support | Private messaging | The relatively high response rate that our intervention providers received shows there is potential to scale direct online engagement across different social media platforms. By deploying intervention providers of different backgrounds and who each used different tones in their delivery, we also gained valuable insights into the types of approaches that were more successful in sustaining response than others. For example, we found a non-judgmental approach in which the intervention provider asked the user to explain or elaborate on their interests in certain extremist narratives, followed by the intervention provider giving a non-judgmental, alternative, or counter-narrative, the best method for starting and sustaining engagement.<br><br>There are several risks inherent in online outreach work that need to be mitigated. These include risks with automation – while automation, or the use of digital tools to filter out potential candidates for engagement and helps scale this work, this may also risk outreach with "false positives," so some form of human review is recommended. Human review is resource-intensive, however, so can prove a barrier if funding is limited.<br><br>Intervention providers may also find it challenging to respond in a timely and effective manner, particularly if they are engaging multiple users and have other commitments. |

| | | | | |
|---|---|---|---|---|
| 30 | ISD's "Reparlons Jihad," "Et toi, le Jihad?," (France, 2017) – an initiative run by a French citizen collective to counter Islamist propaganda, to promote social cohesion, and to encourage the development of critical online engagement. This satirical campaign was developed in partnership with an illustrator and promoted on Facebook. | Counter VE Proposition | Organic reach, Facebook ads generally (type not specified) | 94% increase in Facebook page likes over the three months, 30% increase in engagement (e.g., comments, posts, likes, shares) on the page.<br><br>6,368,441 views of promoted content, 363,888 people directly engaged (e.g., comments, likes, shares) with promoted content.<br><br>8.5% engagement rate (compared with a Facebook average of 5%) and Facebook Relevance Score of 9.6/10 (Facebook provides this qualitative indicator that allows the measurement of the relevance of content posted by pages. This indicator is based on various factors, such as positive reactions (e.g., likes, clicks, video views, etc.) from those that view the content, as well as negative reactions (e.g., clicking on "I do not want to see this content")).<br><br>68% of the audience interacted positively with the content, a favorable sign for the campaign as a whole. This can be confirmed by the messages received via Facebook Messenger, 88% of which were positive.<br><br>The campaign also seems to have played a role in the mobilization of the audience, as many users contacted the page to flag extremist content.<br><br>The campaign's primary content was single-panel comics with a short caption meant to spark debate. Simple, immediate content (including comics, posters, infographics, images, etc.) is a good way of delivering a message to an audience that does not know or trust the campaign, because it requires far less from the audience. Videos are a popular medium and are great for exploring an issue in depth; however, they take time and require an audience to agree to watch it, which is rare in practice. On average, viewers watch less than a third of a video on Facebook and few bother to turn the sound on, especially if watching on a mobile phone in a public place. Immediate content on the other hand delivers the message at a glance and can result in higher rates of engagement because more of those reached will have absorbed the message. Consider how to balance these to increase engagement in campaigns.<br><br>RPJ had good engagement in part because of tight targeting and strategic outreach. When a campaign is working to reach an at-risk audience, popularity can be damaging. "Going viral" may be useful for awareness-raising campaigns or those that wish to promote more general social values, but if a campaign hopes to directly address specific ideologies or belief systems with an at-risk audience, keep the targeting narrow and focus on greater engagement with a smaller number.<br><br>A broad audience can result in a more limited engagement, since it may be more difficult to connect with an audience through more generalized content that has not been tailored for one specific point of view or experience.<br><br>Questions over the effectiveness of targeting every young person in France with anti-extremism messaging: only a tiny fraction of those will be considering extremism. Though for a campaign raising awareness of extremism to a wider audience, this may well be more appropriate.<br><br>Negative reactions represented 8% of all comments and reveal a part of the audience that is opposed to the campaign. While they are never pleasant to deal with, negative responses can actually be an indicator of good targeting and can create opportunities for valuable engagements. When speaking to an at-risk audience with content that directly contradicts the ideology they may already be consuming, one should expect that individuals within this audience will challenge the campaign or defend their own ideas. Reaching these individuals at all means the content reached audiences that may need to see it the most and the interactions with them (if handled well) give the campaign a chance to challenge an adherent of the ideology directly. |

| | | | | |
|---|---|---|---|---|
| 31 | Civic Fab's "What The Fake," (France, 2017) – counter hate, extremism, and manipulation online by checking the hateful speech that proliferates on the internet. The campaign has leveraged factual, verifiable, and well-sourced information, as well as positive content.<br>WTF is active across multiple platforms, including Facebook, Twitter, and YouTube, as well as a dedicated website. | Counter VE Proposition | Organic reach, Unspecified advertising across multiple platforms, including Facebook, Twitter, and YouTube, as well as a dedicated website | WTF succeeded in reaching a broad audience across France. 21,001,780 impressions on Facebook; 9,283,908 individuals reached by promoted campaign materials; 3,720,366 views of promoted videos. Across all videos the average user watched 16.4% of the video<br><br>642,441 engagements with campaign content (3.1% engagement rate), 115,537 page likes. The campaign did exceedingly well in reaching its target audience: reach and impressions were both high and the vast majority of viewers were under the age of 34, with approximately half under the age of 24. Most of the videos had a higher than average retention rate for Facebook, which indicates that the content was appropriate for the target audience.<br><br>WTF is an awareness-raising campaign addressing a social issue that impacts a huge number of people across a population. Therefore, the campaign benefited from a broad target audience that could help improve its reach and help spread its message further afield.<br><br>WTF adopts a conversational tone that is informative, clear, and casual. This tone enables them to engage with a wider audience and helps unify the campaign even as it takes on a wide variety of topics. A conversational tone is useful for informative campaigns geared toward a more general audience because it prevents the content from becoming overly academic or preachy.<br><br>Message: WTF balances content that draws attention to the challenge and debunks other narratives with positive content that offers an alternative. It is important in all counter-narrative campaigns that you do not focus solely on "countering" without giving the audience an alternative to consider.<br><br>Engagement, however, was lower than average, assumed to be in large part due to the nature of the content and the breadth of the targeting.<br><br>Audience: A broad audience can result in more limited engagement, however, since it may be more difficult to connect with an audience through more generalized content that has not been tailored for one specific point of view or experience.<br><br>Also, questions over the effectiveness of targeting every young person in France with anti-extremism messaging: only a tiny fraction of those will be considering extremism. Though for a campaign raising awareness of extremism to a wider audience, this may well be more appropriate. |

| | | | | |
|---|---|---|---|---|
| 32 | Bodyguard AI – Artificial intelligence to moderate comments on social media and flag "toxic content."<br>1. First, the technology looks at the text to detect and clean all emojis, typos, or misspelled words.<br>2. Next, it recognizes any words or groups of words that could potentially be harmful.<br>3. After detecting potentially toxic content, the technology analyzes the context around it, to determine primarily to whom the content is addressed.<br>4. Finally, the technology returns a result after overlaying its analysis with any custom moderation rules set in advance, relating to the severity of the content, toxicity type, and message recipient. | Rebalance perceived cost / reward: Deliver Warning | Scanning content (comments etc.) to detect toxic content | Helps stop the spread of hate speech from a technical point, but doesn't address the person or feelings that generated the hate speech.<br><br>Reliability of categorization is a difficulty. |
| 33 | HateAid, Germany – Their approach is victims-oriented, they empower victims of hate speech/crime online to get back online so that positive voices aren't silenced by extremists. | Provide Direct Support (to victims)<br><br>Build resilience to radicalization | Organic reach, General support for the presence of non-extreme audiences | Interesting indirect way to counter the long-term effects of extremism.<br><br>Hard to quantify effects on extremism, though support for victims is a worthy objective in its own right. |
| 34 | EXIT DE, 'Hass Hilft/Hate Helps', Germany – Racist or extremist comments online are turned into charitable donations. | Rebalance perceived cost / reward<br><br>Counter VE Proposition | Public messaging i.e. counter posting on a comment thread | Raised over 10,000€.<br><br>Difficulty in proving that this helps reduce hateful comments.<br><br>Confrontational nature may risk further alienation. |
| 35 | CST's dontblameme campaign – The campaign, which started on World Peace Day on September 21, 2018, ran throughout the High Holy Day period to promote the joint humanity which all share: that we may categorize ourselves as Muslim, Jewish, LGBT, BAME, or any other individual identity, but we all share our humanity. | Build Resilience to radicalization | Organic reach, some kind of advertising | The campaign, which ran on Twitter, Facebook, and Instagram, reached over 2 million unique users.<br><br>Such a broad audience makes it hard to know whether this peaceful message reached those it needed to (i.e. those with hateful ideas).<br><br>Also unclear how such a popular campaign based on admirable ideals might be received with angry, isolated, unhappy individuals – will it even register? |

| | | | | |
|---|---|---|---|---|
| 36 | ReThink Project – The general objective of the project is to prevent vulnerable audiences from starting a process of radicalization by offering them resilience and critical thinking mechanisms, as well as convince those already engaged within a process of radicalization to abandon it or dissuade them from going further by providing an alternative narrative that deconstructs extremist rhetoric, in order to change violent behavior. | Counter VE proposition<br><br>Build resilience to radicalization | Organic reach, Side-bar targeted ads, Newsfeed targeted ads, Promotion of featured content | Large reach – 5 million+ impressions, 1 million+ minutes viewed, and 1483 comments.<br><br>Story-based content helped engage the audience.<br><br>Danger in placing stats above significance: were all these impressions the vulnerable people that needed to see each campaign's respective message?<br><br>Other platforms – particularly Twitter and Instagram – could have been leveraged more.<br><br>Shorter videos would have helped.<br><br>Facebook policies make it harder to advertise anywhere near topics of extremism, etc.<br><br>Involving influencers could enhance campaigns. |
| 37 | The Misled' rap project to debunk Islamic extremist narratives – The Misled was a multi-platform campaign consisting of a hip-hop album and web-based miniseries developed by Lebanese, Syrian, Palestinian, and Jordan rap artists and creatives involved in research on extremism. The objective was to give a voice to frustrated and disenfranchised youth in the region while demonstrating inconsistencies in Islamist extremist terrorist recruitment messaging. It featured music, a documentary, and a web series. | Counter VE Proposition<br><br>Supply Less Extreme Alternative | Organic reach, presumably unspecified advertising | Garnered more than 150,000 listens and some 1.5 million views.<br><br>Messenger: Using messengers who are able to deliver a campaign's message in an enticing manner that appeals to its audience is vital in generating positive impact. The Misled's use of rap artists and creative content producers who were already knowledgeable of the subject matter enabled the campaign to engage its target audience of Arabic-speaking youth in an engaging way.<br><br>Message Testing: The organizers of this campaign did extensive message testing before publishing the various creative outputs. This phase of the campaign's development incorporated members of the intended audience, whose feedback was vital in refining the campaign so that it would better resonate with its audience.<br><br>Content: Using a variation of media not only helped disseminate The Misled and its content further than a single platform and medium may have but also gave it a greater opportunity for engagement. Members of the audience will have certain types of content and media that resonates stronger with them than others – the use of multimedia allowed the campaign to accommodate to this.<br><br>Qualitative assessment also showed that the campaign positively influenced the target audience, particularly in terms of building resilience to extremist recruitment messaging.<br><br>An integral element in this campaign was the involvement of the target audience in the production of content. Local hip-hop artists and producers, who understood the issues on the ground, created the content. Extensive message testing was conducted to make sure the campaign would resonate with the people it was trying to reach.<br><br>Messengers may lose influence or credibility depending on who the campaign organizer is.<br><br>Difficulty in reach those at risk of extremism specifically, as opposed to those generally disenfranchised. |

| | | | | |
|---|---|---|---|---|
| 38 | HIVE Pakistan – HIVE is a first-of-its kind space in Pakistan, dedicated to training, research, resource development and social innovation to counter extremism and work toward an inclusive, peaceful society. | Build resilience to radicalization<br><br>Raise awareness / skills | Organic reach | Little publicly available evaluation. |
| 39 | Nigeria United – FB/offline campaign in Maiduguri for vulnerable to Boko Haram audience showing Muslim/ Christian cooperation through football and facilitating football activities. | Build Resilience to radicalization<br><br>Supply less extreme alternative | Facebook Page, Ads, Billboards, Event, Radio Roadshow/watch parties | Precise localized targeting and understanding of target audience context (including interplay between ethnic division and interreligious divisions) was important for strategy.<br><br>Leveraging context and target audience's interests (African Cup of Nations football tournament) allowed for a preventative tool to be credible and engaging for the target audience. Linking online and offline communities to build a coherent brand and experience was effective.<br><br>Football is a good vehicle to deliver holistic resilience-building messaging for a young male audience.<br><br>Credibility of messengers (imams, famous footballers, local influencers) was important.<br><br>Smartphone usage and data limitations meant FB videos did not work well; image-based ads and offline event were better. |
| 40 | Ummahsonic – FB magazine brand for upstream vulnerable to Islamist extremism audience promoting shared British Muslim identity. | Build Resilience to radicalization<br><br>Supply less extreme alternative | Facebook ads, Facebook group, Instagram Live, YouTube Ads | Upstream strategy allowed for broad targeting and broad content approach.<br><br>Upstream strategy prevented specific impact measurement as theory of change to radicalization was too distant.<br><br>Need for a lot of day-to-day content to populate channel, therefore expensive.<br><br>Need for very clear attribution policy to avoid negative media coverage, but does that undermine the project? |
| 41 | Open Your Eyes to Hate – counter-hate campaign for vulnerable to far-right audience aiming to raise awareness of and counter extremist narratives. | Counter VE Proposition | Facebook Ads, YouTube Ads, Website, interactive film & paid search | Videos improved over time and some went viral.<br><br>Learned successful formula of short 1-3 minute videos, subtitles, eye-catching openings, and early pivot to main message.<br><br>Paid search campaign performed well and showed impact from the audience (e.g. how do I join the EDL).<br><br>Lose control when a video goes viral, which could lead to unintended audiences coming together in comments, even for polarizing effect. |
| 42 | Great Get Together – annual celebration to unite communities, bridge divides, and tackle loneliness. | Build resilience to radicalization | Event, Website, Facebook Ads, YouTube Ads, Media Articles/Interviews | Important to have mixed online/offline component.<br><br>Drawing on credible figurehead (Jo Cox) and memorable/shocking event (her murder) was important.<br><br>Leveraged positive patriotism to reach diverse audiences (vulnerable to Islamist and far-right). |

| | | | | |
|---|---|---|---|---|
| 43 | Educate Against Hate – provide parents, teachers practical advice on protecting children from extremism and radicalization. | Raise awareness / skills | Website hub, Facebook page (and ads), YouTube page (and ads), Facebook bot pop-up | Official government branded resource built credibility and linked with classroom training.<br><br>Creation of easy-to-navigate hub and synergized web/social presence increased uptake.<br><br>Collaboration with NGOs to boost resources was effective. |

To learn more about the Global Internet Forum to Counter Terrorism (GIFCT), please visit our website or email outreach@gifct.org.