

Introducing 2022 GIFCT Working Group Outputs

GIFCT WORKING GROUPS OUTPUT 2022



GIFCT
Global Internet Forum
to Counter Terrorism

Dr. Erin Saltman
Director of Programming,
GIFCT

In July 2020, GIFCT launched a series of Working Groups to bring together experts from across sectors, geographies, and disciplines to offer advice in specific thematic areas and deliver on targeted, substantive projects to enhance and evolve counterterrorism and counter-extremism efforts online. Participation in Working Groups is voluntary and individuals or NGOs leading Working Group projects and outputs receive funding from GIFCT to help further their group's aims. Participants work with GIFCT to prepare strategic work plans, outline objectives, set goals, identify strategies, produce deliverables, and meet timelines. Working Group outputs are made public on the GIFCT website to benefit the widest community. Each year, after GIFCT's Annual Summit in July, groups are refreshed to update themes, focus areas, and participants.

From August 2021 to July 2022, GIFCT Working Groups focused on the following themes:

- Crisis Response & Incident Protocols
- Positive Interventions & Strategic Communications
- Technical Approaches: Tooling, Algorithms & Artificial Intelligence
- Transparency: Best Practices & Implementation
- Legal Frameworks

A total of 178 participants from 35 countries across six continents were picked to participate in this year's Working Groups. Applications to join groups are open to the public and participants are chosen based on ensuring each group is populated with subject matter experts from across different sectors and geographies, with a range of perspectives to address the topic. Working Group participants in 2021–2022 came from civil society (57%), national and international government bodies (26%), and technology companies (17%).

Participant diversity does not mean that everyone always agrees on approaches. In many cases, the aim is not to force group unanimity, but to find value in highlighting differences of opinion and develop empathy and greater understanding about the various ways that each sector identifies problems and looks to build solutions. At the end of the day, everyone involved in addressing violent extremist exploitation of digital platforms is working toward the same goal: countering terrorism while respecting human rights. The projects presented from this year's Working Groups highlight the many perspectives and approaches necessary to understand and effectively address the ever-evolving counterterrorism and violent extremism efforts in the online space. The following summarizes the thirteen outputs produced by the five Working Groups.

Crisis Response Working Group (CRWG):

The GIFCT Working Group on Crisis Response feeds directly into improving and refining GIFCT's own [Incident Response Framework](#), as well as posing broader questions about the role of law enforcement, tech companies, and wider civil society groups during and in the aftermath of a terrorist or violent extremist attack. CRWG produced three outputs. The largest of the three was an immersive virtual series of Crisis Response Tabletop Exercises, hosted by GIFCT's Director of Technology, Tom Thorley. The aim of the Tabletops was to build on previous Europol and Christchurch Call-led Crisis Response events, with a focus on human rights, internal communications, and external strategic communications in and around crisis scenarios. To share lessons learned and areas for

improvement and refinement, a summary of these cross-sector immersive events is included in the 2022 collection of Working Group papers.

The second output from the CRWG is a paper on the Human Rights Lifecycle of a Terrorist Incident, led by Dr. Farzaneh Badii. This paper discusses how best GIFCT and relevant stakeholders can apply human rights indicators and parameters into crisis response work based on the 2021 GIFCT Human Rights Impact Assessment and UN frameworks. To help practitioners integrate a human rights approach, the output highlights which and whose human rights are impacted during a terrorist incident and the ramifications involved.

The final CRWG output is on Crisis Response Protocols: Mapping & Gap Analysis, led by the New Zealand government in coordination with the wider Christchurch Call to Action. The paper maps crisis response protocols of GIFCT and partnered governments and outlines the role of tech companies and civil society within those protocols. Overall, the output identifies and analyzes the gaps and overlaps of protocols, and provides a set of recommendations for moving forward.

Positive Interventions & Strategic Communications (PIWG):

The Positive Interventions and Strategic Communications Working Group developed two outputs to focus on advancing the prevention and counter-extremism activist space. The first is a paper led by Munir Zamir on Active Strategic Communications: Measuring Impact and Audience Engagement. This analysis highlights tactics and methodologies for turning passive content consumption of campaigns into active engagement online. The analysis tracks a variety of methodologies for yielding more impact-focused measurement and evaluation.

The second paper, led by Kesa White, is on Good Practices, Tools, and Safety Measures for Researchers. This paper discusses approaches and safeguarding mechanisms to ensure best practices online for online researchers and activists in the counterterrorism and counter-extremism sector. Recognizing that researchers and practitioners often put themselves or their target audiences at risk, the paper discusses do-no-harm principles and online tools for safety-by-design methodologies within personal, research, and practitioner online habits.

Technical Approaches Working Group (TAWG):

As the dialogue on algorithms and the nexus with violent extremism has increased in recent years, the Technical Approaches Working Group worked to produce a longer report on Methodologies to Evaluate Content Sharing Algorithms & Processes led by GIFCT's Director of Technology Tom Thorley in collaboration with Emma Llanso and Dr. Chris Meserole. While Year 1 of Working Groups produced a paper identifying the types of algorithms that pose major concerns to the CVE and counterterrorism sector, Year 2 output explores research questions at the intersection of algorithms, users and TVEC, the feasibility of various methodologies and the challenges and debates facing research in this area.

To further this technical work into Year 3, TAWG has worked with GIFCT to release a Research Call

for Proposals funded by GIFCT. This Call for Proposals is on Machine Translation. Specifically, it will allow third parties to develop tooling based on the [gap analysis](#) from last year's TAWG Gap Analysis. Specifically, it seeks to develop a multilingual machine learning system addressing violent extremist contexts.

Transparency Working Group (TWG):

The Transparency Working Group produced two outputs to guide and evolve the conversation about transparency in relation to practitioners, governments, and tech companies. The first output, led by Dr. Joe Whittaker, focuses on researcher transparency in analyzing algorithmic systems. The paper on Recommendation Algorithms and Extremist Content: A Review of Empirical Evidence reviews how researchers have attempted to analyze content-sharing algorithms and indicates suggested best practices for researchers in terms of framing, methodologies, and transparency. It also contains recommendations for sustainable and replicable research.

The second output, led by Dr. Courtney Radsch, reports on Transparency Reporting: Good Practices and Lessons from Global Assessment Frameworks. The paper highlights broader framing for the questions around transparency reporting, the needs of various sectors for transparency, and questions around what meaningful transparency looks like.

The Legal Frameworks Working Group (LFWG):

The Legal Frameworks Working Group produced two complementary outputs.

The first LFWG output is about Privacy and Data Protection/Access led by Dia Kayyali. This White Paper reviews the implications and applications of the EU's Digital Services Act (DSA) and the General Data Protection Regulation (GDPR). This includes case studies on Yemen and Ukraine, a data taxonomy, and legal research on the Stored Communications Act.

The second LFWG output focuses on terrorist definitions and compliments GIFCT's wider Definitional Frameworks and Principles work. This output, led by Dr. Katy Vaughan, is on The Interoperability of Terrorism Definitions. This paper focuses on the interoperability, consistency, and coherence of terrorism definitions across a number of countries, international organizations, and tech platforms. Notably, it highlights legal issues around defining terrorism based largely on government lists and how they are applied online.

Research on Algorithmic Amplification:

Finally, due to the increased concern from governments and human rights networks about the potential link between algorithmic amplification and violent extremist radicalization, GIFCT commissioned Dr. Jazz Rowa to sit across three of GIFCT's Working Groups to develop an extensive paper providing an analytical framework through the lens of human security to better understand the relation between algorithms and processes of radicalization. Dr. Rowa participated in the Transparency, Technical Approaches, and Legal Frameworks Working Groups to gain insight into

the real and perceived threat from algorithmic amplification. This research looks at the contextuality of algorithms, the current public policy environment, and human rights as a cross-cutting issue. In reviewing technical and human processes, she also looks at the potential agency played by algorithms, governments, users, and platforms more broadly to better understand causality.

We at GIFCT hope that these fourteen outputs are of utility to the widest range of international stakeholders possible. While we are an organization that was founded by technology companies to aid the wider tech landscape in preventing terrorist and violent extremist exploitation online, we believe it is only through this multistakeholder approach that we can yield meaningful and long-lasting progress against a constantly evolving adversarial threat.

We look forward to the refreshed Working Groups commencing in September 2022 and remain grateful for all the time and energy given to these efforts by our Working Group participants.

Participant Affiliations in the August 2021 - July 2022 Working Groups:

Tech Sector	Government Sector	Civil Society / Academia / Practitioners	Civil Society / Academia / Practitioners
ActiveFence	Aqaba Process	Access Now	Lowy Institute
Amazon	Association Rwandaise de Défense des Droits de l'Homme	Anti-Defamation League (ADL)	M&C Saatchi World Services Partner
Automatic	Australian Government - Department of Home Affairs	American University	Mnemonic
Checkstep Ltd.	BMI Germany	ARTICLE 19	Moonshot
Dailymotion	Canadian Government	Australian Muslim Advocacy Network (AMAN)	ModusIzad - Centre for applied research on deradicalisation
Discord	Classification Office, New Zealand	Biodiversity Hub International	New America's Open Technology Institute
Dropbox, Inc.	Commonwealth Secretariat	Bonding Beyond Borders	Oxford Internet Institute
ExTrac	Council of Europe, Committee on Counter-Terrorism	Brookings Institution	Partnership for Countering Influence Operations, Carnegie Endowment for International Peace
Facebook	Department of Justice - Ireland	Business for Social Responsibility	Peace Research Institute Frankfurt (PRIF); Germany
JustPaste.it	Department of State - Ireland	Centre for Analysis of the Radical Right (CARR)	PeaceGeeks
Mailchimp	Department of State - USA	Center for Democracy & Technology	Point72.com
MEGA	Department of the Prime Minister and Cabinet (DPMC), New Zealand Government	Center for Media, Data and Society	Polarization and Extremism Research and Innovation Lab (PERIL)
Microsoft	DHS Center for Prevention Programs and Partnerships (CP3)	Centre for Human Rights	Policy Center for the New South (senior fellow)
Pex	European Commission	Centre for International Governance Innovation	Public Safety Canada & Carleton University
Snap Inc.	Europol/EU IRU	Centre for Youth and Criminal Justice (CYCJ) at the University of Strathclyde, Scotland.	Queen's University
Tik Tok	Federal Bureau of Investigation (FBI)	Cognitive Security Information Sharing & Analysis Center	Sada Award, Athar NGO, International Youth Foundation
Tremau	HRH Prince Ghazi Bin Muhammad's Office	Cornell University	Shout Out UK
Twitter	Ministry of Culture, DGMIC - France	CyberPeace Institute	Strategic News Global
You Tube	Ministry of Foreign Affairs - France	Dare to be Grey	S. Rajaratnam School of International Studies, Singapore (RSIS)
	Ministry of Home Affairs (MHA) - Indian Government	Dept of Computer Science, University of Otago	Swansea University
	Ministry of Justice and Security, the Netherlands	Digital Medusa	Tech Against Terrorism
	National Counter Terrorism Authority (NACTA) Pakistan	Edinburgh Law School, The University of Edinburgh	The Alan Turing Institute

	Organisation for Economic Co-operation and Development (OECD)	European Center for Not-for-Profit Law (ECNL)	The Electronic Frontier Foundation
	Office of the Australian eSafety Commissioner (eSafety)	Gillberg Neuropsychiatry Centre, Gothenburg University, Sweden,	The National Consortium for the Study of Terrorism and Responses to Terrorism (START) / University of Maryland
	Organization for Security and Co-operation in Europe (OSCE RFoM)	George Washington University, Program on Extremism	Unity is Strength
	Pôle d'Expertise de la Régulation Numérique (French Government)	Georgetown University	Université de Bretagne occidentale (France)
	North Atlantic Treaty Organization, also called the North Atlantic Alliance (NATO)	Georgia State University	University of Auckland
	Secrétaire général du Comité Interministériel de prévention de la délinquance et de la radicalisation	Global Network on Extremism and Technology (GNET)	University of Groningen
	State Security Service of Georgia	Global Disinformation Index	University of Massachusetts Lowell
	The Royal Hashemite Court/ Jordanian Government	Global Network Initiative (GNI)	University of Oxford
	The Office of Communications (Ofcom), UK	Global Partners Digital	University of Queensland
	UK Home Office	Global Project Against Hate and Extremism	University of Salford, Manchester, England,
	United Nations Counter-terrorism Committee Executive Directorate (CTED)	Groundscout/Resonant Voices Initiative	University of South Wales
	UN, Analytical Support and Sanctions Monitoring Team (I267 Monitoring Team)	Hedayah	University of the West of Scotland
	United Nations Major Group for Children and Youth (UNMGCY)	Human Cognition	Violence Prevention Network
	United States Agency for International Development (USAID)	Institute for Strategic Dialogue	WeCan Africa Initiative & Inspire Africa For Global Impact
		International Centre for Counter-Terrorism	Wikimedia Foundation
		Internet Governance Project, Georgia Institute of Technology	World Jewish Congress
		Islamic Women's Council of New Zealand	XCyber Group
		JOS Project	Yale University, Jackson Institute
		JustPeace Labs	Zinc Network
		Khalifa Ihler Institute	
		KizBasina (Just-a-Girl)	
		Love Frankie	

Active Strategic Communications: Measuring Impact and Audience Engagement

GIFCT Positive Interventions and Strategic Communications Working Group



Introduction

The Global Internet Forum to Counter Terrorism (GIFCT) aims to prevent the proliferation and promotion of terrorist and violent extremist basic content on the internet through collaboration between the tech sector, civil society and governments. Through this strategic and operational partnership, GIFCT works with practitioners, academics and agencies to create work streams and paths to help inform, identify and tackle the multi-layered problem sets presented by harmful actors in the online space. Through its Positive Interventions Working Group (PIWG), an output related to better understanding of active strategic communication from measurement, impact and audience contexts was agreed. The following report provides key learning, barriers and challenges to these issues, as a means to progress this subject area into a more evidence based arena for future efforts. This report focuses its attention on the processes, practices, and challenges of designing, delivering, and measuring online positive interventions within Countering Violent Extremism (CVE) and Counter-Terrorism operational contexts. As such, the report aims to continue (in a developmental manner) the building and sharing of knowledge, practices, and learning that has been led by GIFCT's CAPPI WG on positive interventions. Building on the July 2021 CAPI2 "Positive Interventions" report, which provided a macro-level view of strategy, delivery, and program considerations, this report outlines a more granular and practitioner-oriented effort, focusing on specific elements of the strategic communications process and how this affects and impacts understanding of audience engagement, reach, and measurement. Such an approach offers insights and learning from localized, global, and private-public partnerships that have been created specifically (or in alignment with) CVE needs such as counter-disinformation, harm reduction, critical thinking, and prebunking/inoculation-based prevention work.

There are three main sections of this report that delve into the outputs the CAPPI WG identified as areas of interest: measuring impact or success in campaigns, best practices for audience targeting, and turning passive counter-narratives into active strategic communications. The core content of the report provides examples, conceptual threads, and suggestions for how to make strategic communications more active and grounded through nuanced understandings of audiences, sentiment, and engagement. In doing so, the overall narrative emerging from this effort points to the need for private-public partnerships to evolve in order that they sit firmly at the heart of CVE efforts, driving innovation, trust building, and impact measurements moving forward.

Section 1: Measuring Impact or Success of Campaigns

In the online campaign or project delivery context, the need for measurement and evaluation systems to be part and parcel of the overall design is an accepted norm and practice. Within the commercial arena, this often takes the shape of a set of impact indicators that have become widely known as 'vanity' metrics. The main aim of such indicators is to provide both the delivery agent (advertising agency, government agency/Civil Society Organization (CSO)/NGO) and the 'client' or principal stakeholder with a set of measurements (likes, shares, impressions) and the number of times a piece of content is 'seen' and then 'engaged' further with – known as a Click Through Rate (CTR) – that offer a birds-eye view of the project's outcomes. These metrics also serve as barometers for overall campaign/project success and value for money discussions. However, this practice

requires further development with regards to how such measurements can be viewed within more detailed qualitative and quantitative aspects of behavior change efforts, including greater emphasis on capturing sentiment, longitudinal impact, and the role of other interlocutors (for example the individual's own offline/online eco-system). Within the CVE space, this issue and set of practices take on greater degrees of nuance in regard to how accurately and effectively such metrics can offer meaningful data and results for highly subjective and open-ended issues such as radicalization, deradicalization, disengagement, and desistence. The idea of online positive interventions posits certain assumptions about both the role and potential attitudinal and behavioral impact of the intervention on often hard-to-reach audiences with diverse levels of vulnerability and disparities in their access to information.

Based on GIFCT's CAPPI WG CAPI2 report published in 2021, positive online interventions are operating under the rational belief of the effectiveness of promoting credible, positive alternatives or counter-narratives, while acknowledging that end-users are not simply passive recipients of messaging and the values contained in them. This forms the 'what' element of the project's inception. The second element takes on the need for an overarching goal or objective, which is to counteract possible interest in terrorist or violent extremist groups. This goal also operates under the caveat that interaction and engagement with ideas put forward by positive interventions would be effective and leave an impressive legacy as a result of the experience. Such goals become the 'why' answer for the intervention's need to exist or be initiated. In a related effort, work carried out under the European Commission's Radicalization Awareness Network (RAN) Communications and Narrative WG has initiated and continued to develop the GAMMMA+ model as a practical guide tool for conducting positive interventions within an overarching framework model.¹

A United Nations Development Programme (UNDP) and International Alert collaboration produced a report based on an assessment of a CSO project by PeaceGeeks that lent its focus to different types of proxy variables and attitude types that were possibly indicative of extremist beliefs. Such efforts are seeking to build out and develop a more robust understanding of both the identification and prevention of extremist attitudes.²

Wrapped up within theoretical framework contexts, questions of measurement become aligned to the idea of impact, with the issue of "reach" being viewed as a standard cumulative metric that can offer both evaluative and success criteria outputs for project performance. While the idea of offering reach statistics appears to be logical in terms of wanting to know to who, where, and even how the message, content, or communication was presented, reach can lead to misleading and counter-intuitive sets of metrics without certain key "add-on" features incorporated into the data mining mix. It is even debatable if reach alone can be viewed as equating to impact, and if CVE campaigns should even be aiming for optimum or maximum reach, given the contested nature of radicalization as a concept and the dangers of mass audience targeting in this context. In essence, when assessing

.....
 1 "RAN C&N Effective Narratives: Updating the GAMMMA+ model, Brussels 14-15 November 2019," European Commission: Migration and Home Affairs, n.d., https://ec.europa.eu/home-affairs/pages/page/ran-cn-effective-narratives-updating-gamma-model-brussels-14-15-november-2019_en.

2 "Design, Monitoring & Evaluation of PVE Projects in Jordan - A Baseline Assessment of PeaceGeeks' Projects," International Alert, 2022, <http://www.pvetoolkit.org/design-monitoring-pve-project-in-jordan>.

reach metrics in the CVE online interventions space, certain unique factors and conceptual issues affect project performance in ways that can skew both what a project has set out to achieve and the results that are gathered from vanity metrics and reach measurements. These warrant further discussion in order to ensure a more effective relationship between project design and project outcome. The key to online CVE interventions rests with the planning and insight phase, as these two issues directly affect audience targeting, project measurement, and performance results, as well as creating a baseline assessment through which impact is ultimately understood and evidenced later on.

In order for CVE online interventions to move beyond reach statistics, the design of such efforts should prioritize incentivizing engagement upon being reached instead of merely reach statistics. What matters is the nature of the engagement and it yielding positive outcomes for deradicalization /disengagement efforts. Reach data alone can be too broad to provide specificity in counter-radicalization efforts. For example, one of the first questions that needs to be addressed is what the evaluation needs actually are? Evaluating projects in order to prove the efficacy of a policy stance, such as the UK government's efforts to counter extremism through its "Prevent" agenda, has been seen as a "scatter-shot" approach designed to produce 'quick wins' because of a lack of perceived risk and proportionality rationale. Evaluations can also be conducted to justify the project's existence beyond the 'why' logic framework, but these typically rely on highly subjective and contested ideas such as values, citizenship, and certain definitions of what 'moderate' versus 'extreme' views may mean.

This is evident in the UK government's social cohesion and counter-extremism approach ("Contest Strategy"), which has often conflated promoting cohesion to mean the reduction of extremism by using behavioral science-based evaluations in such a broad context as to undermine producing reliable data/outcomes. Although this example is from a policy perspective, the operational issues of conflating 'mainstream' social cohesion efforts with countering extremism often stem from having too broad an approach to defining problem sets and targeting audiences without adequate insight or rationale. The issues of seeking to increase cohesion levels and reducing the influence of extremism have now started to be separated through more distinct lines of effort. For example, GIFCT's theoretical framework approach (mentioned above) differentiates between resilience building within a prevention context (further from harm) and undermining extremist ideology in more upstream counter-extremism contexts (further towards harm). By way of example, the UK Home Office's "Building a Stronger Britain Together" initiative via the work of its Research Information and Communications Unit (RICU) offers detailed insight into these issues.³

An even more pertinent strand of this issue would be to ascertain if the evaluation is needed in order to measure or present an empirical analysis of successful deradicalization or disengagement. In this context, the type of campaign and its tactics (e.g., resilience building, capacity building, support services, deradicalization, or counter-radicalization) should ultimately determine the nature and role

.....
 3 Home Office, "Evaluation of the Building a Stronger Britain Together (BSBT) programme," GOV.UK, July 29, 2021, <https://www.gov.uk/government/publications/evaluation-of-the-building-a-safer-britain-together-bsbt-programme>.

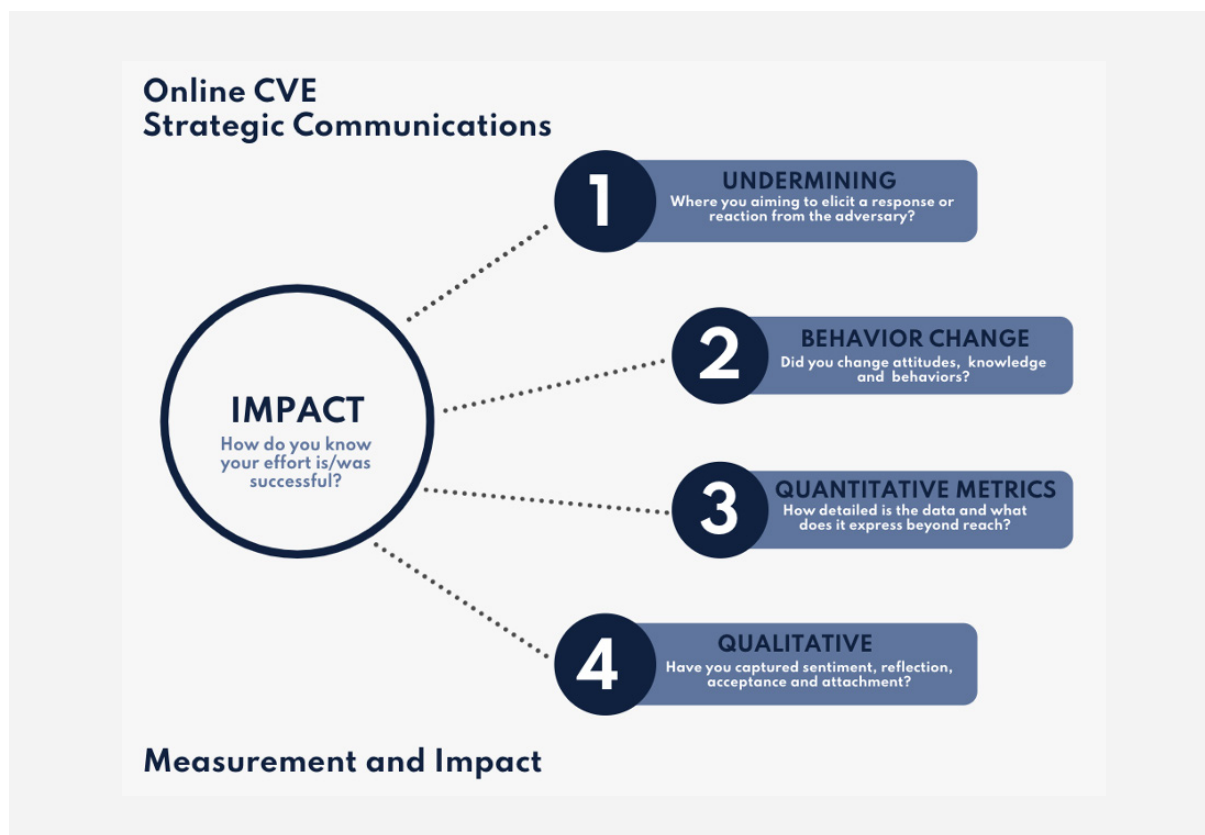
of the evaluation approach as well as its suitability in achieving aims and objectives.⁴ There is little doubt that reach statistics need to provide many stakeholders and interested parties with 'catch-all' data sets that can serve different purposes depending on the line of inquiry. Both offline and online interventions in this space are unable to claim any sole causal link to someone's life journey, choices, and influences over time, as they are based purely on the 'impact' of the intervention. As such, no measurement and evaluation model or framework can make any similar claim. Therefore, the goal of such work is based on the mitigation and prevention of harm at the deterrence level alongside challenging and exposing harmful narratives at the undermining level of such interventions. Such efforts are designed to reduce the likelihood of harm and seek to protect those who may be most vulnerable to narratives and potential actions. The "measurement of success debate" is one that should be seen as working towards creating better practices through a greater understanding of what works and what doesn't in specific contexts and environments. The better the specific measurable outcomes, the greater the likelihood of the overall online CVE interventions space being more and more effective.

The aim of isolating and identifying true causal links to certain decisions or actions/ behaviors is complicated by the fact that online identity is couched inside numerous "influence" parameters that are impossible to capture and codify. This is even more nuanced in the realm of CVE and it is of great importance that the strategic framework created for an online intervention must offer an exceptional level of understanding of the problem set, insight into audience types and lifestyles, as well as key expertise in the problem set being explored (be it deradicalization or capacity/resilience building). Simply knowing the type of intervention required at the strategic level (deter, redirect, etc.) does not address the depth of insight needed to deliver the intervention with an effective plan and model. This applies equally to the issues of reach and ensuring the "right" audience(s) have been reached. This is where the strategy and design phase are so crucial in regards to the direct relationship that exists between insight and audience – namely the more robust the insight, the more accurate the audience targeting.

While it is acknowledged that the size, scale, and context of the protagonists involved in this space create a huge variance in regard to capacity and resourcing matters, it must also be acknowledged that good planning and strategy will fundamentally improve the prospects of a campaign being "successful" or "effective." Therefore, certain useful practices can be put in place (e.g., top-level strategy, theory of change, effects framework, adversary analysis, audience segmentation matrix) to ensure that design, delivery, and measurement are not only aligned with each other but also with the overall objective of the project or campaign in question. Some of the key considerations in regards to reach and impact can be addressed by ensuring that a clear set of objectives, indicators, and data collection processes are in place, as per below:

.....
 4 "CTED Analytical Brief Countering Terrorist Narratives Online and Offline," United Nations: Security Council - Counter-Terrorism Committee (CTC), 2020, <https://www.un.org/securitycouncil/ctc/content/cted-analytical-brief-%E2%80%93-countering-terrorist-narratives-online-and-offline>.

Figure 1: Measurement and Impact Considerations



Strategy and Objectives

Questions and considerations relating to impact and reach are best managed through the development of an effective strategy. Does the strategy and desired end-state align with the objective(s) set out by the stakeholders, problem set, or client? In this context, the term strategy refers to the logical framework, proposition, and rationale created to express what needs to be done, why it needs to be done, and how it will be done. A key area that is often overlooked in CVE online intervention planning is the issue of the effects (both intended and unintended) of the campaign, message, or content being part of the public domain. Strategy and its efficacy form the core component of what needs to be done and how this will be done:

Figure 2: Strategy Considerations



A practice that can to some extent help to alleviate over-reliance on generic reach data is the use of an “effects framework,” which is essentially a set of outcome indicators at a granular level that can identify and measure the impact of key tactical methods, content, or messaging deployed by a campaign as part of an overall theory of change. As the effects framework needs to take shape during the strategy and planning phase, its outcomes are tied to both strategic intent, objectives, and impact. This comes down to a simple proposition: if, for example, behavior change is one of the key aims of a project, one of the objectives must be something similar to “reducing the impact” or “undermining the narrative” of the antagonist/adversary. The reason for this assertion is obvious in the sense that any intended target audience would need to change its behavior towards a Violent Extremist Organization (VEO) through actions, which would then need to be picked up by the effects framework. This change (ideally) needs to have been planned for, which is why it is part of the effects framework, and as a result should also be able to be identified, measured, or verified. For an online intervention, reach statistics alone cannot sufficiently determine this change or variance, so more nuanced impact and effects measurements are required. The role of strategy in communications in this regard cannot be underestimated.⁵

.....
⁵ Haseeb Tariq, “Five Components Of A Successful Strategic Communications Plan,” Forbes, June 22, 2021, <https://www.forbes.com/sites/forbes-communicationscouncil/2021/06/22/five-components-of-a-successful-strategic-communications-plan/?sh=326789e05813>.

Going Beyond Reach

Classically, calls to action were seen as useful indicators in assessing the extent to which audiences engaged with a message, understood its premise, and then actioned desirable efforts through active choice and decision making. Advances in social media platforms and tech tools combined with greater degrees of audience lifestyle segmentation have meant that relying solely on online reach data can create misleading and skewed notions of both audiences and their needs. A simple example would be to view an individual's reaction to a piece of content (like, dislike, share, retweet, etc.) as a barometer of their values set, preferences, or position on a subject matter. This also applies to contextual disparities that exist, because one piece of content delivered a certain way may have entirely different sentiment values attached to it by the same person in different circumstances. Such issues are pertinent not just to reach and impact matters, but equally to ensure the differences between extremism and violent extremism.

The “golden goose” of online CVE interventions remains the relationship between content and engagement, followed by values alignment and behavior change. The idea that genuine and often challenging engagement lives beyond generic reach metrics should drive the design and narrative elements of the positive intervention. The notion of active communication is one that essentially elicits a response beyond pressing ‘like’ and takes both the content and its end-user into a new space of engagement that may be challenging or affirming but allows the end-user to express ideas and sentiments and (ideally) make positive choices. Of course, the context of the campaign largely dictates the type of engagement that may be possible (e.g., building followers and generic metrics versus supporting a local network to improve governance structures to undermine extremist influences), as deradicalization efforts are markedly different from resilience building efforts in terms of tone, message, and objectives as well as the type of action or desirable behavior sought. However, engagement is also important in and of itself, because comments and other types of interactions give other users social clues about the content.⁶ Comments can actually change users' perceptions of the video/article they are seeing, so community engagement and management can not only bring important benefits for evaluation⁷ but actually support the impact of campaigns.⁸

An area of effort that is gathering pace in regards to both approach and content design is “prebunking”-based interventions. Also known as work that takes an inoculation theory approach as its central proposition, moving firmly into the preventative space with interventions that see “prevention, not cure” as the best way to engage in both the CVE and disinformation/misinformation space, prebunking efforts have made significant strides in both design and capability. A central development in this form of approach has been to move away from the “debunking” tactic into the building of “mental armor” through prebunking content that is designed to prepare someone to identify, assess, and make informed choices when extremist or disinformation content appears on their screen or social media feed. However,

6 Franklin Waddell, “What Does the Crowd Think?” *New Media & Society* 20, no. 8 (August 2018): 3068–83. <https://doi.org/10.1177%2F1461444817742905>.

7 Jae Eun Chung, “Peer Influence of Online Comments in Newspapers: Applying Social Norms and the Social Identification Model of Deindividuation Effects (SIDE),” *Social Science Computer Review* 37, no. 4 (August 2019): 551–67. <https://doi.org/10.1177/0894439318779000>.

8 Hue Dong, Hong Vu, and Long TV Nguyen, “Effects of Online Comments on Risk Perception,” *The 70th Annual Conference of International Communication Association*, Gold Coast, Australia, January, 2020. https://www.researchgate.net/publication/338751162_Effects_of_Online_Comments_on_Risk_Perception_and_Intention_to_Communicate.

there are challenges with this approach, specifically how adding significant size and scale would be managed and still retain elements of nuance and stratification. Future development of this approach can seek to further refine this element. The work of Kurt Braddock provides an excellent roadmap into the emergence of prebunking/ inoculation efforts into the CVE terrain.⁹ In collaboration with Kurt Braddock and others, Jigsaw (a division of Google) has attempted to advance the practical application of debunking efforts and offered both testing phase results as well as setting out more contextual criteria for how this work can continue to develop.¹⁰

Debunking-based efforts can sometimes struggle to match both the quantity and propensity of extremist content simply because their dissemination does not have a pace sufficient enough to challenge and then move end-users' attention span away from the original harmful content. Some remnants, ideas, or narrative elements remain, and this is why the building of mental, emotional, and digital resilience through prebunking is seen as a more effective method in preventative spaces. The logic behind this approach is simply that if enough examples of extremist-based content are shown to users through prebunking tactics, they will be better equipped to identify and question it.¹¹

Measurement and Evaluation

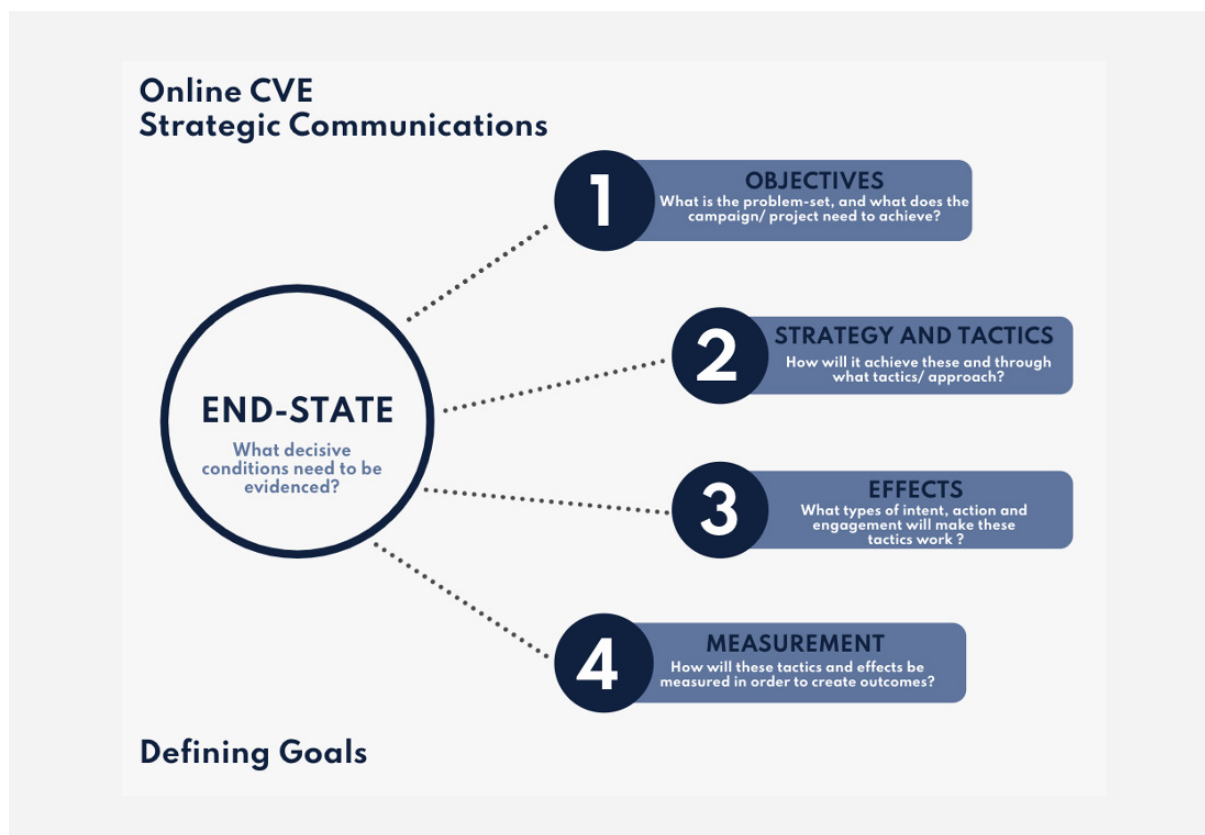
In order to enhance and improve the quality of outcome indicators that a project may be seeking to identify through its content dissemination, the assessment framework requires two core pieces of information at the outset: the desired end-state and the means to have even a simple baseline understanding of the intended audience's level of support, sentiment, grievance, hope, fear, etc., of the issue being tackled. The baseline assessment provides the most robust and reliable method of identifying how much movement is needed during the lifespan of the campaign from the audience and how this can be effectively measured (which should already be incorporated into an effects framework). This baseline assessment also allows the desired end-state to be measured so as to ascertain how well its conditions have been met. If such processes and practices are not in place, there is a strong likelihood that results will not go beyond vanity metrics and birds-eye view statistics that still plague the online CVE space. This highlights the need to have a clear end-state goal in mind in order to be able to measure outcomes and impact appropriately:

.....
 9 Kurt Braddock, "Vaccinating Against Hate: Using Attitudinal Inoculation to Confer Resistance to Persuasion by Extremist Propaganda," *Terrorism and Political Violence* 34, no. 2 (2022): 240-262, doi: 10.1080/09546553.2019.1693370.

10 Kurt Braddock et al., "Engagement in subversive online activity predicts susceptibility to persuasion by far-right extremist propaganda," *New Media & Society*, (February 2022), <https://doi.org/10.1177%2F14614448221077286>.

11 "Psychological Inoculation: New Techniques for Fighting Online Extremism," Jigsaw, June 24, 2021, <https://medium.com/jigsaw/psychological-inoculation-new-techniques-for-fighting-online-extremism-b156e439af23>.

Figure 3: Defining Goals



Sentiment and Engagement

The ideal scenario within an online CVE intervention is that it has some synergy with efforts in the offline space. The idea of an “audience” obscures the reality of terms like ‘network,’ ‘community,’ and ‘activism’ at local levels. Online audiences are mysterious and hard to “read” in the isolation of an offline reality, which exacerbates the issue of some audiences being “hard-to-reach.” A way around this quandary is to attempt to base much of a project’s content focus on engagement with people offline before they become audiences. However, how much or how little offline engagement may or may not be required is dependent on the campaign’s goals and tactics. This approach is also linked to a need for the CVE efforts in the online space to move beyond the traditional focus group route and start to incorporate audiences throughout the inception, dissemination, and evaluation stages. Although resources for CSOs/NGOs are limited, these entities technically still have better access to community networks and actors who help bridge the gap between the project and the experience of the local population. Engagement built from open and trusting practice has a better chance of succeeding than unverifiable campaigns that can arouse suspicion. The closer the offline engagement is to the online iteration, the greater the chance of success for the project over time. This is also true of sentiment and tracking sentiment change. The need to understand the concept of the “audience” in the online space cannot be understated and is very much at the heart of successful

planning for sensitive campaigns.¹²

Meeting Communities Offline

In ideal circumstances, the measurement and impact needs of a project or campaign benefit greatly from practitioners having access and engagement with audiences/ communities in offline settings. In this context a key advantage to delivering positive interventions from a community/CSO/NGO perspective is that of physical access to various potential target audiences. By the nature of their community origins and placement, such organizations are in a position to engage in a more hands-on fashion with networks, interest groups, and individuals to advance project aims and objectives. Traditionally, this means that levels of trust and mutual support offer projects the chance to be more grounded in real-world needs and also allow for greater flexibility in using creative and diverse means and mediums to deliver targets. The approach used by PeaceGeeks (<https://peacegeeks.org>) to this issue is a very useful place to begin understanding the offline/online nexus. Community engagement that is grounded in offline relationships allows design and planning to go beyond the norm of “focus group” testing with relatively unknown audiences to a process that can add diversity to different possible target audience needs with bespoke roundtable and workshop style events that allow greater nuance and depth to be added to project design elements. This results in delivering content that is to some extent the product of a partnership effort between the CSO/NGO and the community, with the latter taking on the role of stakeholder in a basic form. This type of community engagement can add significant value to the intervention’s credibility and longevity among online audiences. Although the CSO and private sector do not always naturally fit together in certain contexts where nuances may need to go beyond “cultural expertise” from a distance and instead be grounded in the real-world experience of vulnerable communities, in terms of the need for efficacy in approach and tactics, the commercial sector possesses several useful avenues for CSOs to explore to enhance their existing offline engagement.¹³

Testing Reactions to Content

A project’s chances of creating a successful impact are closely linked to how effectively the testing or pilot phase is at identifying strengths, weaknesses, challenges, and nuances. The initial phase of such testing often involves focus group discussions on a basic concept or idea, which allows both prospective audience members and the project team to understand positive and risky or more negative elements of the design. This can be further enhanced through simple variations to the ‘A-B’ testing model, such as rendering a few different versions of the same core content and disseminating them to similar target audiences to test and gauge reactions to the subtle variations in content and its presentation. The Abdullah-X project, which was a CSO/YouTube partnership, used A-B testing through creating different thumbnails with varying degrees of intrigue and then targeted different potential users through hashtag/metadata changes with the same core content.¹⁴ A-B testing clearly

.....
12 Molly Riddle-Nunn, “Tips and Tools for Understanding Your Online Audience,” Mostly Serious, June 22, 2018, <https://www.mostlyserious.io/news-updates/understanding-your-online-audience>.

13 Aritya Sen, “7 Offline Customer Engagement Strategies You Didn’t Know About,” Involve Me, September 14, 2020, <https://www.involve.me/blog/offline-customer-engagement-strategies/>.

14 Abdullah-X, “Freedom of Speech vs. Responsibility,” YouTube.com, March, 2014, <https://www.youtube.com/user/abdullahx>.

offers results and data on what content elicits greater engagement, but this is not as easily applied to impact measurement. Another way to elicit more nuanced testing phase feedback is through an iterative approach that is open to comments and suggestions emanating from the sample audience in regards to possible language and tone as well as cultural and related nuances.

There is an argument that testing using more extreme or polarizing content can exacerbate existing conditions. Despite this potential risk, this approach also serves as a means to maintain an element of tolerance towards changes needed to content in order to build or maintain user engagement and trust. Such an approach offers what is known as a 'network effect' to come into play; audiences that are familiar with the content, having engaged and been part of a testing phase, can subsequently amplify the reach of the end-content to their online and offline networks.

The Abdullah-X project underwent rigorous testing, feedback, and adjustments from all perspectives, including branding, content style, duration, narrative, sentiment analysis, and audience targeting. The project also engaged with potential target audiences offline, taking both the content and concept into classrooms across the UK, enabling audiences to help shape the overall design and glean information on elements that offered more engagement and curiosity. This process employed A-B testing, where more counter-narrative/challenging versions of content were tested for different thumbnails, hashtags, and metadata to ascertain if more radicalized individuals or ISIS supporters were more or less likely to click depending on what was used.

Translating Offline Efforts into Online Domains

A key way in which offline efforts can form an important aspect of online intervention approaches is through adaptive delivery, which is different from reactive delivery by virtue of its content tone and form. In the adaptive context, project content design retains an element of scope to re-render and re-order narrative tone and form to more closely match the needs of target audiences through engagement analysis. Where content is based on social media posts or audio programming (for example radio/ podcasting), the tolerance for amendments and subtle changes is clearly easier. For more creative delivery such as video, animation, serials, etc., the key learning from the offline domain is to ensure that the initial testing phase can glean feedback on both concept, creative treatment, narrative, and tone.¹⁵

A key aspect of the offline experience is the individual user experience that can still be found within a group or community delivery context. This translates into the online setting through content that can house intellectual, emotional, and personal "touch points" from one overarching delivery strand. Examples where this method is implied within the project parameters can be found in the One2One¹⁶ campaign and search redirect intervention approach.¹⁷

.....
15 Andrew Glazzard, "Losing the Plot: Narrative, Counter-Narrative and Violent Extremism," ICCT, May 22, 2017, <https://icct.nl/publication/losing-the-plot-narrative-counter-narrative-and-violent-extremism/>.

16 Ross Frenett and Moli Dow, "One to One Online Interventions – A Pilot CVE Methodology," ISD, September, 2015, <https://www.isdglobal.org/isd-publications/one-to-one-online-interventions-a-pilot-cve-methodology/>.

17 "Redirect Method Canada: Final Report," Moonshot CVE, March, 2021, <https://moonshotteam.com/resource/canada-redirect-final-report/>.

When a campaign is seeking to engage mass audiences, narrative form and content style play a large part in either being able to create effects at the individual level through critical reflection or using persuasive messaging to encourage individual actions that help achieve bigger end-state objectives. Both of these types of effects are in their way linked to the idea of content being able to access and influence the cognitive space of the end-user, seeking what is commonly known as a “cognitive opening.”¹⁸

Another aspect of translating offline methods to the online domain is found in the management of negative feedback, pushback, or “blowback” of content from audiences who may or may not be part of an intended target audience. Often seen as a gray area of the CVE world in terms of a campaign’s remit or scope to address such issues, this scenario is often found where campaigns gain the type of trust or credibility usually reserved for face-to-face offline interventions in the deradicalization and prevention space. Dependent on whether a campaign’s focus is providing alternative narratives or some form of counter-narrative, negative feedback allows a window of opportunity to directly address grievances, misinformation, or hate speech through tact, poise, and intellectual prowess.

These are attributes not usually associated with extremist narratives, which are generally known to be binary and obstructive in nature and tone. This does not, however, mean that extremist narrative is not complex in narrative, form, and tone. The issue is that pushback means the campaign’s overarching presence has caused a reaction from those who are clearly part of the problem. An appropriate well-crafted response is a very effective way to undermine this feedback and build presence in an already congested online space. The opposite applies to rushed, panicked, and not well thought out responses, which may inadvertently amplify the extremist messaging. Ultimately, the scope and context of a project will determine if this approach is viable or not, but the fact remains that direct engagement of negative feedback that is well-constructed and tactical in nature can provide a campaign with enhanced trust and behavior change possibilities. Both audience feedback and behavior change issues fall neatly into the next section, which looks at how engagement is understood through audience sentiment broadly.

Different Approaches to Measuring Sentiment Change

The role and scope of communications-based efforts to counter and challenge issues like extremism remain limited without other inputs being in place at a larger scale. While this can be construed as a loss-leader if taken literally, the power of communication lies in its ability to apply and interpret both its presence and the presence of communication originating from other actors. In the context of positive interventions, this applies to the form, tone, quality, and quantity of communication received and analyzed by a campaign from its target audiences. This process is commonly known as sentiment analysis – the process of detecting positive or negative sentiment in text. The focus of this effort is on the identification of certain types of ‘polarity’ that can be inferred from within the study of text (e.g., positive, negative, neutral).

.....
 18 First introduced in Quintan Wiktorowicz, *Radical Islam Rising: Muslim Extremism in the West*, (Lanham, Md: Rowman and Littlefield Publishers, Inc, 2005).

Sentiment analysis as a tool, however, goes further than the relatively basic process of polarity detection into areas of analysis that interpret specific feelings and emotions (e.g., happy, sad, angry), urgency, and even intentions (e.g., interested versus not interested). These can be further branched into distinct processes that seek to grade different sentiments against certain desirable and undesirable criteria, engage in emotion detection analysis from rudimentary text, perform aspect-based analysis of sentiment driven by specific factors or conditions, and multilingual sentiment analysis that looks at variations of the same sentiment in different language contexts. Key benefits of sentiment analysis as an approach include its ability to sort data at scale, real-time analysis, and employing consistent criteria from which to analyze. However, there are methodological issues with how this approach would be used in prevention spaces, as there are challenges with inferring tone with written words, especially in massive data sets and across cultural boundaries. Polarity is also a challenge with mid-polar terms that rely on tone, sarcasm, culture, idioms, and context. Sentiment analysis is a useful tool for judging individual or community valence when there are drastic shifts in polarity, but more often than not, it is not reliable at scale.

It is important to note that sentiment and effectiveness are not synonymous, particularly when they relate to interventions in the realm of violent extremism. One of the most common validity issues with measures of intervention effectiveness is the assumption that when participants “feel” something, it is evidence that the intervention is effective. That is not necessarily true. Discrete emotion theory tells us that different sentiments have different action tendencies, meaning that when an individual experiences a particular emotion, she is motivated to act in a specific way. For example, if an individual experiences anger or disgust in response to an intervention, she is likely to discount or mentally dismiss the intervention. In this way, it is critical to understand (a) the emotions that intervention participants experience, and (b) their action tendencies in association with those emotions. The lesson here is that all emotional responses should not be considered equivalent and that making this assumption can lead to counterproductive intervention design.

Following from this, the measurement of discrete emotions (or sentiment) can be undertaken in one of two ways. First, thought-listing protocols can be very useful. In a thought-listing exercise, participants are simply asked to “free-write” about their perceptions of the intervention. Typically, participants are provided about 10 minutes to do so. There are no rules or restrictions as to what participants can write. Following this thought-listing exercise, two or three coders look at all the ideas written by participants and code them. Although coding schemes can be designed differently, at a minimum they should include two measures: (1) ‘relevant’/‘not relevant’ and (2) ‘positive’/ ‘negative.’ Ideas coded as ‘relevant’ and ‘positive’ can be considered affirmative sentimental responses to the intervention. A second method of measuring sentiment in relation to an intervention is a traditional survey. Surveys can be designed in a number of ways, but a recommendation is to embed sentiments of interest in a larger emotional scale.

Generally speaking, measures of beliefs and attitudes in response to intervention practices need to be tailored to the intervention so participants can be asked specifically about elements of the intervention. For example, if the purpose of the intervention is to reduce conspiratorial beliefs, belief scales should ask about those specific beliefs; if the purpose is rather to affect attitudes in relation to a specific minority group, then the attitude scale should be tailored to measure whether participants

have positive or negative impressions of that group.

A theory relevant to measuring beliefs, attitudes, and sentiment is Reasoned Action Theory. The premise of this theory is that behaviors stem from beliefs about a behavior, norms surrounding the behavior, and one's capacity for engaging in the behavior. In other words, exposure to certain messages can affect beliefs about violence, which affects attitudes about violence, which affects intentions to engage in violence, which in turn affects violent activity.

Perhaps the most sought-after aspect of project measurement and evaluation is the need for analysis that considers diverse touch points or nodes of narrative engagement and sentiment. This is where more advanced forms of sentiment analysis through the presentation of visualizations come to the fore. Analysis can consider overall sentiment, sentiment over time, sentiment by rating, and sentiment by topic. By putting these touch points together or as a suite of measurement tools, different effects and sub-effects can be identified and correlated to glean if a certain narrative has met its desired conditions or objectives. This approach also serves as a very effective way to measure change in different contexts and conditions. Such a process has fast become the norm for online interventions that rely solely on the power of communication to create an effect. A key tool in sentiment analysis is the manner in which tech tools can be used to "opinion mine" in ways similar to how data mining has emerged as a standard practice for large data sets. The opinion mining process standardizes the analysis element in order to measure "impact" or results, but is not without its drawbacks, some of which can affect the overall success criteria of a project/campaign or specific narratives.

The challenges of a sentiment analysis approach are diverse and carry significant risks that challenge both CVE and counter-disinformation efforts in similar ways. Subjectivity and tone are two particularly difficult issues to determine when analyzing complex data sets from audiences who may have very different ways in which they culturally or religiously express their views and may require the need to localize certain ideas. In terms of polarity identification, issues relating to contextual realities must also be considered. Factors such as the use and subsequent detection/analysis of irony, humor, and sarcasm are still areas that require more effort and expertise in capturing and interpreting. The way comparisons are made and their criteria can lead to controversy when it comes to reliable measurements of sentiment. The interpretation and understanding attributed to the use of emojis/icons is also an underdeveloped area of this field and requires further work.

Some of the most pressing issues involve how a sentiment analysis framework defines 'neutral,' which can be seen as a relatively subjective choice. This also links to the issue of human annotator accuracy versus machine learning systems using algorithms to both detect and interpret data. Despite these drawbacks, sentiment analysis has many potential avenues to follow with respect to the issues of interpretation and measurement of sentiment change, which are still key indicators of the possibility of behavior change. It is also noteworthy to point out that extremists in various guises have become more and more adept at harnessing sentiment analysis for their own narrative propagation needs. In the next section, a more nuanced discussion of these possibilities is introduced.

An area of the reach and impact debate that is full of rich potential is ensuring that measurement

of sentiment change considers fully the distinctive nature of sentiment that is culturally specific, driven by more individualistic tones, and its origins and manifestations in ideological underpinnings. When these three nodes align within the expression of (for example) one person, the sentiment change analysis requires both flexibility and a degree of sophistication to interpret language, form, and tone in an equally nuanced way to its originator's levels. This can also be applied to sentiment that is more difficult to clearly categorize in terms of positive or negative, etc., due to it posing a set of complex assertions that may fall in between affirmative, rejectionist, challenging, or questioning tones of language/sentiment. This is particularly relevant to complex interventions with audiences who may be sympathetic to elements of extremist thinking (conservative or literalist thinking) and apathetic to other elements (worldview or violence). Such complexity brings into focus the issues related to subjectivity, the need for weight to be given to the contribution of indigenous studies in CVE efforts, and the relative differences between various groups in terms of power and influence or representation even.

A way through these challenges is to apply the study of semiotics alongside sentiment analysis for an even richer picture. Semiotics offers a study of signs and symbols that locate their meanings and values within specific normative and cultural contexts.¹⁹ This enables certain language and expression motifs (colors, words, terms, icons, and symbols) to be attributed interpretive values based on the cultural conditions that create the value in the first place through shared experiences, stories, and communities. A classic example of this would be how the term 'conservative' holds many meanings in different cultural, religious, and political contexts, so it cannot be given an overriding interpretive meaning without risk. When combined, certain colors and symbols create different meanings in different conditions (such as flags and slogans) and may serve very different purposes as a result. Both design and planning as well as evaluation and measurement stages can take advantage of semiotics to create a more in-depth awareness of possible outcomes and what these may mean. The more culturally competent the design, the more likelihood there is of impact and sentiment change being valid and able to direct future needs and interventions.

This all takes the issue of sentiment change and its analysis into a more comprehensive and defensible space and allows narratives to be more effective in the operating environment and not just in the online domain. The key to this process is seeing sentiment and change as phasic inputs created during the lifespan of a project's cycle. Capturing sentiment change is a process and sentiment analysis as well as semiotics are tools to aid this endeavor.

Section 2: Best Practices for Audience Targeting

Audience targeting, both from a conceptual and practical perspective is an effort always in evolution and transition based on available tools, knowledge and access. However, certain general understandings of how best to approach do exist, in the shape of moral, ethical, transparency and sensitivity/safety needs. First and foremost, a human rights and rule of law lens should be at the core of any audience targeting. Such processes and understanding naturally vary, depending on who the actor is and the area of operations chosen for the positive intervention. From a CSO/

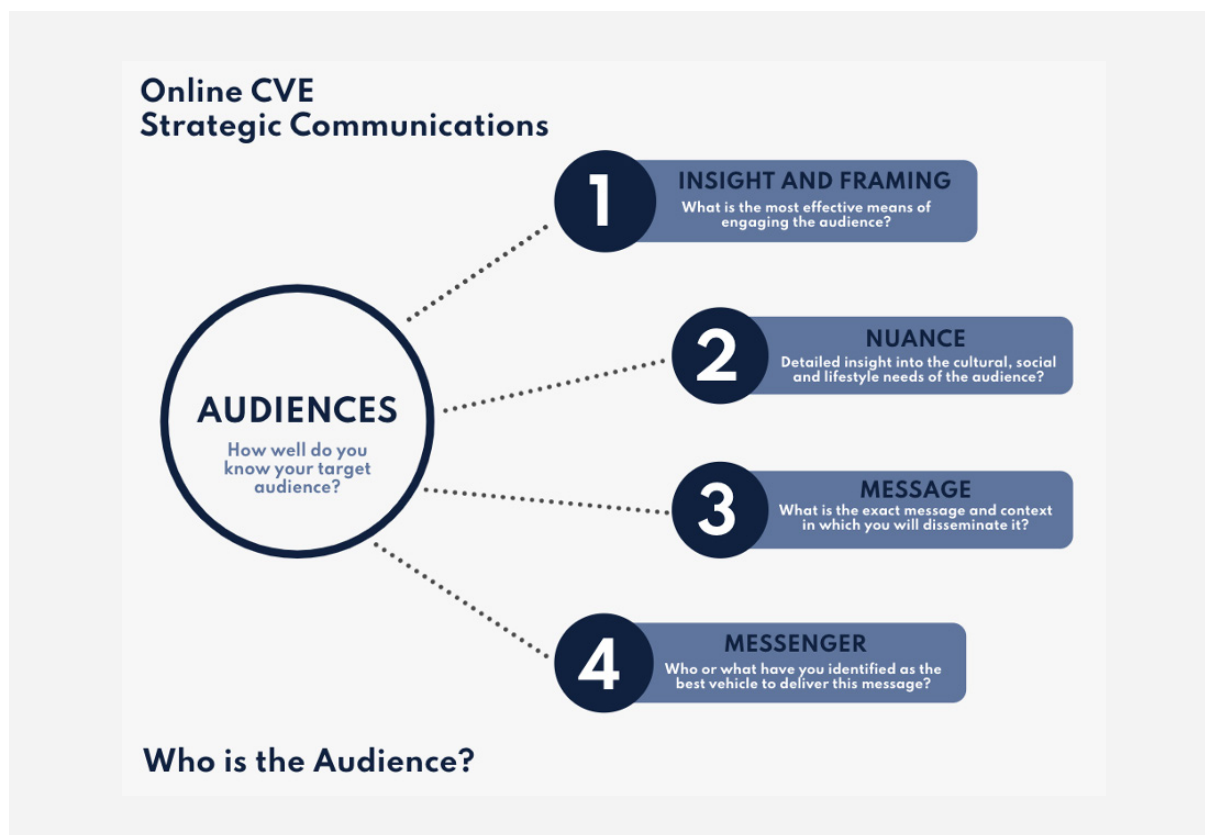
.....
 19 "What is Semiotics?," Sign Salad, 2021, <https://signsalad.com/our-thoughts/what-is-semiotics/>.

NGO perspective, this process should be intertwined closely with the offline relationships/networks that are already in place or can be used to create new networks for specific project needs. In such contexts, the building of trust, meaningful engagement, and mutually beneficial relationships with offline communities allow the CSO sector to not only understand community needs but also translate these into tailored audience segments for the purposes of communication efforts.

From a planning perspective, audience targeting needs to retain a crucial position within related and specific aspects of project design. The basics of strategic communication theory began from a simple yet effective trident of “what, to whom, how” while nowadays there is added emphasis on “where.” The question of “to whom” represents the all-important need for identifying, testing, and confirming the right audience has been found for the message to be disseminated to (and not always in a geo-location sense, hence the rise of “where”). With more and more technological tools at the disposal of many different actors in this space and the rise of the use of information as a weapon in influence and disinformation, it can be argued that audience targeting is now fast becoming the key to success for interventions based on communication.

The backdrop to this emergence can be sourced from the manner in which digital identity and lifestyle choices have given classical understandings of preference and choice a radically new sense of meaning and direction. Multiple-platform identities, online shopping and consumerism, online community networks, and the use of gamification techniques in different contexts have all helped to shape a more flexible understanding of what ‘audience and taste’ may mean. This increased flexibility offers the individual massive choice, but this equally creates additional layers of nuance for those entities seeking to attract, target and engage these audiences. Regrettably, QAnon has been particularly adept at using gamification to attract audiences, and this suggests innovation must continue to drive CVE development needs. In essence, the idea of ‘an’ audience is now something that needs various caveats attached to it to enable more robust and confident targeting, as merely assuming homogenous audiences still exist would be a mistake. Despite this increasing complexity, audience targeting still hinges on two foundational elements being conducted effectively: identification of a problem set and a robust manner through which insight informs strategy. Audience targeting and insight require detailed processes of research, engagement, and segmentation in order to offer the messaging the best chance of engagement and impact.

Figure 4: Audience Targeting



For a positive intervention to have a chance at a relevant impact, the correct weight to the identification of the problem set must be added. Too often, problem sets become rudimentary and basic issues that are based on 'go-to' terms and industry understanding, rather than clear, robust, and defensible decisions regarding why a need for an intervention exists. In terms of mass audience campaigns, terms like "resilience building" or "increasing civic participation" have important rationales behind them, but when used in smaller and more tailored areas of operations, these terms represent different interpretive and assumptive meanings. In conflict zones or areas of instability, resilience building can mean being able to increase community support networks for greater self-reliance, whereas in other contexts it means the creation of networks or even building greater levels of critical thinking.

This applies equally to assumptions made about audiences within the academic field that may (for example) misconstrue vulnerability to extremism by suggesting environment or ideology as being a more relevant cause for fostering extremism or radicalization, whereas in actual real-world contexts environment and ideology are very difficult to separate as singular causal factors. The debate between Olivier Roy and Giles Keppel is a good example of this kind of discourse.²⁰ In choosing only environment or ideology to base both audience targeting and project objective setting, messaging, credibility, and impact can be adversely affected. 'Insight' and 'audience' are therefore not catch-all

.....
²⁰ Adam Nossiter, "That Ignoramus: 2 French Scholars of Radical Islam Turn Bitter Rivals," New York Times, July 12, 2016, <https://www.nytimes.com/2016/07/13/world/europe/france-radical-islam.html>.

terms. Insight into what issues, on whom, and with which tools (e.g., tech, subject matter expertise, partnerships) involve scoping questions that require holistic treatment to be able to create the depth of information from which to put a strategy in place for a campaign.

Another useful way to approach audience targeting is to reverse the research parameters in order to formulate ideal and worst-case scenarios for the intervention life cycle. A small community-led intervention should consider broader, more macro audiences as ideal types in terms of desirable conditions, whereas a larger campaign with significant resources can consider creating clusters of micro-audiences who may have some key overlaps in terms of need and behavioral patterns that can help to better create and refine messaging through nuanced narrative and tone. This approach is a variation of the traditional use of A-B testing model but adapted to offer audience targeting a more robust and nuanced approach. With the rise of a more flexible and multi-layered form of online identity coming to the fore, granular level audience segmentation efforts and more nuanced audience insight processes enable projects to refine, identify, cross-reference, and target more relevant audiences through eliminating assumptions and relying on data driven feedback. This could be as simple as adding 'pattern of life' metrics into audience targeting, that capture lifestyle, network affiliation, tribal affiliation etc, in order delve deeper into possible entry points for various audiences who may share some similar factors, but are divergent on others.

As a result, being able to plan content and narrative themes can be processes developed with much greater levels of confidence when information mining lies at the heart of both conceptual and tactical considerations. In a traditional sense, demographics include age, gender, ethnicity, and location. These metrics offer scope to target certain audiences in certain places who may share similar features. If, however, lifestyle choices and network affiliation are considered, new layers and levels of nuance from 'similar' groups emerge that can radically alter both targeting and messaging needs. In the context of CVE efforts in the online domain, it is rarely considered that those deemed to be more 'vulnerable' to extremist messaging or ideology are also in today's world the same people who have a multitude of different identity facets that can be accessed as means to address alternative or counter-narratives.

This is where tech tools can be such powerful and transformative assets for campaigns. Examples would include insights that capture creative lifestyles, entertainment and leisure, and networking affiliations. When information mining reveals relevant and intriguing engagement possibilities, the issue of providing content that may trigger the possibility of a cognitive opening on issues linked to extremism or harmful ideas has more subtle entry points through which to build engagement, trust, and longevity. However, the management and use of information mining techniques and practices raise ethical and risk-based issues that should not be overlooked. Such approaches continue the trajectory of interventions having more nuance and adaptive thinking behind them in order to achieve outcomes that are behavior-change oriented. Essentially, in this way both the planning and delivery phases of a campaign can take a more considered approach to how knowledge, attitudes, and behaviors are initially understood and then shaped through content. Although the ways in which the commercial campaigns sector approaches these issues may differ slightly, the fact that they operate under the same assumptions suggests there are tips and tricks to be learned from the

approaches of other sectors.²¹

Ethical Issues Around Targeting Vulnerable Audiences

The advent of the European Union's General Data Protection Regulation legislation has given greater layers of safety to online users and given the commercial sector much to consider in terms of how audience insight and targeting can now be pursued from legal, moral and ethical standpoints. Historically, micro-targeting capabilities were able to use mobile/digital advertising criteria that allowed for highly specific forms of audience preference to be created via web-advertising and 'ad-displays'. For the commercial sector, this is now even more advanced via micro-targeting tools that use third-party data mining capabilities to be able to tie smartphone data (cell numbers and handsets), IP data, and geo-location tagging in order to offer highly targeted user profiles where required. Seen as standard practice in the commercial sector and the means by which the world's biggest corporations are able to attract customers they believe will be repeat, loyal and long-term prospects, this technology can easily transfer to the online intervention space for strategic communication needs. However, some ethical issues emerge in this context and questions of transparency, attribution, and intent permeate the technology versus engagement versus objectives debate.

This is of particular relevance when the issue of vulnerability among individuals and groups (and its many facets) is considered. One of the first factors that warrant attention is how the term vulnerability is used within the online intervention space, considering the improbability of any actor possessing a complete and holistic assessment of the level of vulnerability present in a group or person. Another consideration stems from how data mining, storage, handling, and aggregation issues coalesce with working in and around vulnerable groups/ audiences. The need for robust quality assurance, security protocols, and risk management systems to be in place is without doubt a core need in this regard, alongside proportionality and transparency.

Vulnerability is an equally descriptive and relative term, and as such is often used to make grand assumptions about target audiences without a comprehensive understanding of how risk and protective factors affect people in similar socio-economic circumstances in different and divergent ways. In areas of conflict and suffering, this issue is still one that cannot be taken at "face value" when content and narrative are the only tools to achieve positive change. Project planning for online positive interventions needs to consider how to first mitigate against any known vulnerabilities through adequate risk management systems, potentially offering signposting and safeguarding support that is either directly or indirectly woven into content design and narratives as well as building tolerance within the project's scope to engage with any emerging risks or vulnerabilities that may arise once the content is live and being engaged with. Addressing vulnerable groups online is linked directly to real-world vulnerabilities that nevertheless manifest differently in each case. A project plan should include an impact assessment phase at both the initiation and end-stage, while also collecting midway point information on how content, audiences, and wider issues have collided

.....
 21 John Lee, "5 Practical Tips to Step Up Your Audience Targeting Game," Search Engine Journal, July 2, 2020, <https://www.searchenginejournal.com/5-practical-tips-to-step-up-your-audience-targeting-game/373428/#close>.

or combined to either increase or reduce known vulnerability levels.²²

Use of Influencers as Credible Messengers for Audiences

The issue of working (or seeking to work) with credible messengers has traditionally been based on access, levels of trust, and risk versus benefit considerations. The rise of the term ‘influencer’ has given the issue of credibility greater bandwidth, but not without some associated issues. The central one is that attempting to define an influencer is still very much an area housed within the commercial marketing sector, and variances of meaning regarding this term are numerous. A broad understanding of the term suggests an influencer is someone who is able to get relatively large numbers of followers to follow through with purchasing or supporting a specific product, service, or cause.²³ Within the CVE context, credibility has morphed to mean different things, given that in specific areas of work deradicalization and hate speech could mean vastly different things. An example would be the use of former extremists (“formers”) to help deradicalize individuals, as per the UK governments Channel program, versus using an NGO/CSO to communicate a message to counter hate speech by virtue of its community access translating to mean ‘credibility.’ It is also important to acknowledge the scarcity of openly available tools for measuring the specific impact an influencer or credible messenger may have had within a campaign. This issue needs addressing if the use and role of influencers is to become a data and research-driven practice within CVE for all actors within the field.

In an era where emerging social and fringe movements are hard to differentiate from each other, extremism and grievance have become more entrenched inside projectionism and “cancel culture” under the umbrella of the so-called “post-truth” era. This is perhaps best demonstrated when the terms activist, activism, and active participation are reviewed in the guise of online communication, engagement, and action. This applies to how differentiation between extreme views does or does not correlate to violent extremism and where the nexus points exist for policy and strategy considerations at governmental and commercial levels. If the rise of influencer culture is considered against these terms and contexts, it has to be taken as accepted that an influencer is given this moniker because there are identifiable and known metrics on this person’s social network presence and reach. The term influence is closely allied to the term persuasion and the basis for much commercial, strategic, and ad-hoc communication is clearly to achieve some sort of persuasion-based impact, whether through subtle or overt influence regarding knowledge, attitude, or behaviors. If the lessons gleaned from audience targeting are taken into account, influencers are most influential when they are able to persuade others to make certain choices, think certain things and behave in certain desirable ways. Provocation strategies employed by digital influencers could therefore be closely studied and then adapted for the purposes of online CVE efforts.

What qualifies someone to be called an influencer in the CVE space is a subject that requires more research and understanding. This can range from being a former, an expert, academic, fashion

22 Jess McBeath, “Supporting Vulnerable Groups Online,” UK Safer Internet Centre, 2022, <https://saferinternet.org.uk/guide-and-resource/supporting-vulnerable-groups-online>.

23 Jenn Chen, “What is Influencer Marketing: How to Develop Your Strategy,” Sprout Social, September 17, 2020, <https://sproutsocial.com/insights/influencer-marketing/>.

expert, sports person, etc., and is of course entirely dependent on the context. In the CVE space, what isn't up for debate is the idea that a genuine influencer is someone who is able to exert some sway over the ideological impact, environmental conditions, and cognitive capacity of a stranger online who, regardless of having some link to the influencer, may have several other more relevant influencers and influences impacting their daily opinion-forming and action-taking. Influencers are worthy of consideration if they are demonstrably able to exert measurable influence over the choices of others, which can be traced and measured to confirm the causal relationship.

Depending on the objectives of the campaign, the level of influence is a key factor in looking and harnessing influencer support for a project. There are degrees and variations of influence at work in the CVE space, which means the choice of influencer is a crucial one, simply because a level of influence does not constitute credibility automatically in the complex environment of ideology, narrative, groupthink, and psychological motivations. The more nuances the era of digital identity has erected means influence and credibility are not immediately aligned features of any campaign seeking to undermine an extremist narrative or movement. The wrong influencer can create more vulnerability and take audiences further from safety, and the notion of the credible messenger has also become too much of a catch-all term when choice and preference have become so expansive, dynamic, and disposable. For example, a religious conservative influencer in practice may have been seen as a way into influencing vulnerable groups away from violent extremism but can equally be far removed from the values and norms of a certain society in regards to respect for rule of law, freedom of speech, and democratic processes. Another even more significant issue is how much influence fear and insecurity play in how audiences take in or ignore the overtures of "credible" messengers or influencers, where VEOs and insurgencies hold sway over daily life. From a community context outside of the CVE space, what 'credible' means for the role of the credible messenger²⁴ can be an interesting contrast to the CVE-specific perspective.²⁵

This leads to another important facet of this issue, namely the assumptions behind the term credible. Actors can attribute the label of credibility to those who may have a certain persona online or reputation but lack the requisite coverage offline, and therefore can be seen as somewhat of a trojan horse to suspicious audiences. The different social worlds from where actors/campaign creators and audiences often come from mean that vanity metrics and persona play a bigger role in assuming a person's credibility than their network coverage and links to different audiences on complex topics. This can therefore create a "catch-22" situation once content is live and linked to a certain person/influencer. The ideal scenario is one where the project team takes on the responsibility for a process of localized due diligence regarding a holistic assessment of the suitability of potential influencers and credible voices, taking into consideration both risk and benefit scenarios. Two areas within this issue that warrant closer attention are when youth audiences are involved and also the issue of working or enlisting "formers" to assist a campaign in a CVE context.

In a predominantly youth engagement context, two factors for consideration come immediately

.....

24 "A Transforming Approach to Justice," Credible Messenger Justice Centre, n.d., <https://cmjcenter.org/approach/>.

25 Michael Jones, "Through the Looking Glass: Assessing the Evidence Base for P/CVE Communications," Royal United Services Institute, July 17, 2020, <https://rusi.org/explore-our-research/publications/occasional-papers/through-looking-glass-assessing-evidence-base-pcve-communications>.

to the fore; what the term youth implies relative to age and how safeguarding and support (including peer support) are either incorporated or offered through third parties. In the online positive intervention context, the usual approach has been to have a lenient stance on age limits in regard to how a campaign or project determines its audience and participants for any workshops, focus groups, or peer-to-peer work. Youth can be defined in such contexts as anyone from 18 to 35 years of age. In many respects, this approach is logical given how the understanding of radicalization and vulnerability to it have fostered significant preventative efforts both online and offline. When looking at recruitment patterns of VEOs, this age range accounts for the vast majority of recruitment data available in the public sphere. It is also interesting to note that in other strategic communication arenas, where a specific cadre of VEOs or one main VEO is occupying territorial, social, and political influence, the same age range is known or referred to as military age males (MAMs).

What this contrast in terminology does is highlight that in most cases within the CVE sphere, youth are the key target audience, and therefore it becomes necessary to ensure that project parameters and pathways to engagement have fully understood and administered safeguarding and support inputs to protect against project, audience, and reputational risk, as well as impact assessments of content that is eventually present in the same environment as narratives of extremism, etc. This is where the “do no harm” principle comes into its own, by means of planning and programming that seeks to mitigate risk through identification, assessment, actions, and management as well as promote positive behaviors through socio-cultural norms and values. It is often possible to develop and support peer-led networks in the offline realm to facilitate such efforts and over recent years this trend has also made its way into online interventions that access digitally connected user communities to teach them how to self-manage and develop narratives through a “network effect.”

In the context of using influencers and credible voices in a youth-focused project or campaign, attention must be paid to the viral nature of social media influence, content, and attention/retention of interest. With the rise of platforms like TikTok, attention spans, optimum content style and length, remixed content, and repurposed content have all been moved forward into new common practices. This evolution has impacted the idea of influence, influencers, and reach in different ways. Where once positive interventions sought to retain audience engagement over longer periods of time, advances in technology and content creation now mean that bite-size content, both static and animated/ video, often has a three-to-ten second impression window and a thirty-to-sixty second duration at maximum. This gives rise to the proposition that the nature and tone of counter-narratives, alternative narratives, and strategic communication in general need to evolve in terms of their intended reach, impact, and desired end-state. Such a change would require a rethink in terms of storytelling, creating emotion and challenging/undermining harmful narratives in shorter and shorter spaces of time, when youth audiences in particular are of importance to a project's needs. In this respect, learning and adopting lessons from the commercial sector offer NGO/CSOs an effective way to leverage these new approaches to positive effect. This is why the more private-public partnerships that are based on equitable and transparent relationships, the more effective and innovation-driven CVE interventions will become. This applies to support and technical needs in equal measure, as matters such as content and its optimization are changing at a rapid pace.²⁶ The issue

.....
 26 Claire Beveridge, “11 Easy Social Media Optimization Techniques to Start Now,” Hootsuite, January 13, 2022, <https://blog.hootsuite.com/social-media-optimization/>.

of how trends can shape and affect design and dissemination can equally not be understated or overlooked in this context.²⁷

In regards to the use of formers as credible voices or messengers, much learning and outcomes (but not necessarily impact) driven information already exists. One2One, AVE Network, Abdullah-X, and Extreme Dialogue offer various important reference points for projects and organizations working in the CVE space. Some of the key factors in this context are that just by virtue of being a former, this does not necessarily mean the individual has knowledge, know-how, and tact in the realms of creativity, safeguarding needs, private-public partnership, and the role of tech. In many ways, this places even more responsibility on the project to create a conducive environment for the former to be able to “add value” to the narrative and for the audience to be able to resonate as well as relate to the presence of the former and their message. This again links to the increasing need for the private sector to assist with training, capacity building, and industry-standard tools to help formers become more accessible and properly utilized in order to leverage credibility, collaboration, and the continuation of initiatives.

Relatability stems from common themes or experiences, whereas resonance implies that a connection goes beyond its original purpose and influences other areas of interest or people of interest. For a former to be ultimately effective as a carrier of a positive message, tone, context, and attribution issues require careful consideration. This is because a brilliant counter-narrative can be self-sabotaged by it being attributed to parties and actors that cause suspicion within the audience, thus causing the former to lose face and the narrative to be attacked. The Abdullah-X project met with this exact fate as a result of enormous media attention, being paraded around the global CVE circuit and eventually offering jihadist sympathizers the chance to frame it is a Western attempt to “undermine Islam.” It became relatively easy to track the projects’ media presence and tie it to organizations and individuals of interest for its adversaries. Language and dialect, cultural nuance, and visual presentation all play a role in how credible a message is and how credibly a messenger is received, even if this happens to be a former. With rapid changes in engagement and reach related issues, it may be useful for positive interventions to be planned with the idea of credibility only being able to be understood once the audience has decided how credible a person or message really is to them and for how long. The credibility issue has several other considerations within sectors such as capacity building, advocacy, and offering support services.²⁸ In regards to accessing and working with victims of terrorism, other considerations and sensitivities²⁹ come into play and require both management and tact to be able to activate and manage such relationships.³⁰

Creating a Trustworthy Online Presence

When engaging in digital work, especially when the aim is to establish direct contact between

.....

27 “Social Trends 2022,” Hootsuite, 2022, <https://www.hootsuite.com/research/social-trends>.

28 Marina Tapley and Gordon Clubb, “The Role of Formers in Countering Violent Extremism,” ICCT, April 12, 2019, <https://icct.nl/publication/the-role-of-formers-in-countering-violent-extremism/>

29 “Counterspeech: Extreme Lives,” Facebook, n.d., <https://counterspeech.fb.com/en/initiatives/extreme-lives/>.

30 Guillaume Deniox de Saint Marc et al., “Handbook: Voices of victims of terrorism,” Radicalization Awareness Network, May, 2016, https://ec.europa.eu/home-affairs/system/files/2020-09/ran_vvt_handbook_may_2016_en.pdf.

an at-risk or radicalized person and a counselor/P/CVE professional, questions of credibility and trust building are closely linked to transparency. This means any professional wanting to engage in this field needs to invest time and thought into creating a convincing online profile that allows insights into their personality and background without including private information to maintain safety. A well thought through profile may be able to create a sense of intimacy and trust that can be able to bridge the distance even in digital settings. Professionals also need to be truthful in their communication about their professional affiliation and objectives. According to many practitioner experts, using anonymous profiles and false pretexts to start conversations with hard-to-reach target groups has proven to be unreliable and unsuccessful. Danish expert Christian Mogensen from the Center for Digital Youth Work has previously spoken about his visible track-record of publications on positive aspects of gaming and online cultures, which, according to him, increased the initial trust of “Incel” community members and enabled him to engage them. While this level of online exposure may not be possible/wanted for regular practitioners, the organization or program with which practitioners are affiliated needs to ensure that their communication around their work supports trust building prior to engaging with a project. The strategy on how a program or practitioner presents themselves needs to be adapted based on the specific topic area/phenomenon they are working on.

Facilitating Cognitive Openings

When looking at the core aims of any strategic communication effort, it goes without saying that some form of behavior change comes high on the list of desired end-states or goals. The shape and form of this behavior change is a different matter, and for this reason, when considering the need to impact audiences at a behavioral level, some form of knowledge and attitude building or shaping needs to occur in order for the behavior change process to be monitored and captured. This process comes together under the guise of creating a cognitive opening. In contrast to positive interventions seeking to create some form of cognitive opening in the opposite direction, extremist narratives can create ruptures in the cognitive alignment of individuals, resulting in cognitive dissonance (when existing beliefs and newly formed extremist beliefs clash and create inconsistency between belief and overt behavior or action). A cognitive opening seeks to redress inconsistencies and create a new window of thought, conceiving different scenarios and applying reason to emotions and narrative cues. Within the positive intervention context, there are numerous ways to seek this type of opening and various tactics that align themselves with innovation and creativity to achieve the desired outcome, which is to take audiences further away from harmful ideas and actions. In section 3 of this report, further exploration on the various ways and means of seeking a cognitive opening are discussed in order to build upon the conceptual and practical implications of this type of work.³¹

Incentivization

A relatively obvious means to seek a different cognitive path for an individual or group is to offer some form of incentive or reward for certain types of behavior or actions. This can be through increased civic engagement at the local level, competitions, prizes, innovation labs and tech-led challenges, or tangible routes to manifest aspiration levels. The general aim of such an approach is

.....

31 Casal Bertoa and Jose Rama, “Polarization: What Do We Know and What Can We Do About It?,” *Frontiers*, June 30, 2021, <https://www.frontiersin.org/articles/10.3389/fpos.2021.687695/full>.

to provide a stimulus of motivation, confidence, and self-reliance that ideally can combine to create a greater sense of personal agency. An incentive-based approach can be used in many of the standard positive intervention objectives already mentioned (e.g., deter, intervene, prevent, build and empower) and is perhaps most effective when resilience building and capacity building aims are at the forefront of the project/campaign's desired outcomes.³² Another example of a delivery model in this context is to use peer-to-peer efforts, which can unite engagement, empowerment, and localized efforts.³³ PeaceGeeks' project with the Meshkat community in Jordan also offers a highly localized and community-focused approach to tackling hard-to-reach groups and addressing sensitive topics.³⁴

Faith as a Facilitator of Cognitive Openings

Using ideologically-based narratives for introducing influence via non-traditional or non-institutional religious beliefs and ideas is a more complex process in the positive intervention space given issues with representation and matters of authority and authenticity alongside accuracy of messaging in regards to meaning and interpretation. Having said this, much of the apprehension around ideologically-based messaging or content stems from a lack of contextual knowledge as to how extremist narratives misuse faith for sinister means and ends. It is very difficult to challenge harmful narratives that include distorted renderings of religious text if the response is not created in a clear, credible, and meaningful context and in a compelling and creative manner. There have been several effective attempts to harness ideologically-based narratives that have accessed figures of religious authority, historical narratives, creativity, rhetoric, satire, and critical thinking tools to deliver predominantly religious messaging aimed at undermining extremist ideology and narrative. Such examples tend to suggest that elements of religious understanding can be packaged in various narrative forms to encourage self-reflection and questioning, and as a result some can contribute to positive cognitive openings.

Two important yet little-known examples of when faith-based narratives opened avenues for the non-violent rendering of certain textual sources are the Mardin declaration of 2010 which addressed and refuted disinformation efforts with some significant faith-based authorities.³⁵ Similarly, the Amman message convened two hundred Islamic scholars from around the world with the aim of differentiating mainstream Islamic principles from extremist renderings of the faith on core subject areas that are often exploited and intentionally misused.³⁶

However, the issue remains that even such important examples as those noted above have not been repacked and repurposed in creative online ways to have a bigger impact. This perhaps speaks of a lack of awareness and skill set in some areas of CVE work to adequately understand and then

32 "Counterspeech: Resiliency Initiative," Facebook, n.d., <https://counterspeech.fb.com/en/initiatives/resiliency-initiative/>.

33 "Counterspeech: Peer-2-Peer," Facebook, n.d., <https://counterspeech.fb.com/en/initiatives/p2p-facebook-global/>.

34 "Meshkat Community Amplifies Arab Voices for Social Inclusion in response to online hate, violence, and polarization," PeaceGeeks, n.d., <https://peacegeeks.org/meshkat-community>.

35 Shaykh Abd al-Wahhab al-Turayri, "The Mardin Conference – Understanding Ibn Taymiyyah's Fatwa," Muslim Matters, June 29, 2010, <https://muslimmatters.org/2010/06/29/the-mardin-conference---a-detailed-account/>.

36 "The Amman Message," n.d., <https://ammanmessage.com>.

juxtapose the nexus between faith, motivation, and intent in many extremist ideas and narratives as well the challenge of being able to bring nuance to harmful viral ideas. This issue is even more complicated when the online context is the terrain in question. Issues relating to trust building and specific skill sets linked to digital literacy and practitioners' understanding of these (e.g., informed choice making, source verification, online safety, etc.), are important issues for the CVE community as a whole to continue to develop in the hope of bridging gaps. More work is needed to bring better faith-based understanding and mainstream narratives from within faith contexts into online CVE planning and delivery. Lack of specialized religious or cultural understanding can no longer be seen as an excuse for poorly designed ideas that can have adverse effects on faith-based communities. A role for the research and tech community could emerge as a keeper of repositories of up-to-date learning and resources from such contexts in partnership with viable CSOs.

Lack of Digital and Social Media Literacy Among “Traditional” Practitioners

It remains a major challenge that a significant portion of highly experienced practitioners, especially from the secondary and tertiary prevention levels, are themselves not experienced online like their digital native target groups. Tech companies could invest in training practitioners to improve their capabilities within the digital world and support P/CVE efforts at all levels. Additionally, practitioners require insights into the latest developments of extremist narratives and activities online so they can act and react in a timely manner. And even beyond that, they need to be aware of the key topics and trends that currently move their target groups. There is a lack of concise practice-oriented analysis regarding such developments that practitioners can refer to during their daily work – ideally with a country- or even region-specific component. Most practitioner organizations do not have the resources to do this themselves. A good practice in this area is the continuous basis monitoring implemented by Modus|Zad in Germany (funded by the German Federal Agency for Civic Education).³⁷ The project produces short monthly/quarterly practice-oriented reports based on a previous analysis of relevant extremist actors online. As of 2022, additional dissemination tools (such as webinars and workshops) are being implemented.

Use of Humor

Humor can be a very effective tool for engaging audiences, building momentum in user following and reach, and communicating complex issues in accessible and memorable ways. Its usage and efficacy depend largely on the ability of content creators to form and maintain “touchpoints” of interest and then curiosity through a medium mainly used for entertainment. Humor that is delivered through comedic and satirical mediums does need appropriate sensitivity checks to mitigate against negative reactions and the risk that messaging is taken out of context and reused by extremists or harmful actors to attack the very auspices of counter-extremism efforts.

Use of memes has proven to work well with some target groups. They provide a low-threshold opportunity to touch upon difficult topics and to elicit reactions and a debate from the target group, which can lead to further engagement and sometimes even direct contact and counseling. When

.....
 37 “Basis Monitoring (2021/22),” Modus | Zad, n.d., <https://modus-zad.de/schwerpunkte/monitorings-trendanalysen/basis-monitoring-2021-22/>.

thematizing difficult and potentially disruptive topics which are bound to raise emotions among the target group in a non-humorous way online, reactions often include considerable adversity, including hate messages and threats. Generally, however, whenever working with humorous or disruptive content, practitioners need to plan for time to observe and engage in the debates that arise as a result. Careful planning is definitely necessary regarding human resources, capability and capacity. There are examples of the use of comedy and humor within P/CVE adjacent fields.³⁸

There is clearly scope for using comedy, satire, and humor within online and offline CVE efforts,³⁹ but it must also be acknowledged that such tactics carry risks, which need to be managed, mitigated, and accepted with a degree of caution and consciousness.⁴⁰

Crisis Points and Deployable Content

Crisis communications require speed and often pre-prepared messaging to be combined with context specific needs. The link between this need and audience targeting issues is based around the nature of the crisis and who the first-hand affected parties are, the larger community around them and then the national/ international context. Based on the emergency, defining who to communicate to is to some extent driven by who is most affected and then working backwards in order as suggested. In a moment of crisis, communication becomes a key interlocutor for both government agencies, communities, and mainstream media. The need for timely, accurate, and trusted accessible information becomes the primary objective in such situations. For those involved in managing and disseminating crises communications, other variables of equal importance are also part of the strategic framework that forms effective “crisis comms.” A key need is to be reactive in a manner that is appropriate to the actor’s position within the overall context. For example, a government agency may be required to provide timely and accurate information relating to emergency response needs, emergency numbers for people to call, information as to what citizens should and should not do, where to go and where to avoid going, etc. The tone and choice of language and messaging style needs to be informative and avoid undertones that could exacerbate already tense issues further. It is important to acknowledge if information is simply not known.

The response to the 2019 terror attack in Christchurch, New Zealand remains both a best practice example for crisis communications, but also an example of how to communicate information without compromising the integrity of the office of government, the needs of the grieving and concerned families, and being steadfast in naming what the act is and isn’t in order to get ahead in the war of narrative and information race that sadly commences in such instances. This example perhaps best demonstrates the value of quick, clear, and robust communication needs in crises situations, and how this can galvanize the community in a positive direction. A key example how addressing engagement

.....
38 Shezo Media, Facebook, n.d., <https://www.facebook.com/ShezoMedia/videos/>; Fata Byyano, Facebook, n.d., <https://www.facebook.com/Fatabyyano>; PRODUCTIONS ON OUR PALETTE, Kharabeesh, n.d., <https://www.kharabeesh.com/entertainment.php>.

39 Inari Sakki and Jari Martikainen. “Mobilizing Collective Hatred Through Humor: Affective–discursive production and reception of populist rhetoric,” *British Journal of Social Psychology* 60, no. 2 (April 2021): 610–634, <https://bpspsychub.onlinelibrary.wiley.com/doi/full/10.1111/bjso.12419>; Olga Khazan, “The Dark Psychology of Being a Good Comedian,” *The Atlantic*, February 27, 2014, <https://www.theatlantic.com/health/archive/2014/02/the-dark-psychology-of-being-a-good-comedian/284104/>.

40 Tom Dreisbach, “How Extremists Weaponize Irony to Spread Hate,” *NPR*, April 26, 2021, <https://www.npr.org/2021/04/26/990274685/how-extremists-weaponize-irony-to-spread-hate?t=1651651473874&t=1652439765302>.

and communication in points of crisis, where sentiment and reaction are very sensitive issues for all involved, was the “Je Suis Charlie” effort in the aftermath of the terrorist attack in France.⁴¹ In the immediate days and subsequent outpouring of emotion and need for healing and grief, coverage of the campaign was very much supported and seen as a necessary statement of unity and strength in a time of loss and anger.⁴² Similarly, in Nigeria the “Bring Back Our Girls” campaign achieved significant global coverage and ensured the actions of Boko Haram were contextualized from the victims’ perspective rather than fuel more coverage for the group’s aims.⁴³ The transference of sentiment and support made this campaign an effective use of social media and network effects at a time when confusion, anger, and fear had gripped the community affected.

Other actors (Government agencies) may have roles that require their communication to be of a reassuring nature, or the need to reaffirm certain core values to worried citizens. Messaging that sends out a clear affirmation of a willingness to challenge, undermine and defeat violent extremism may well be the domain of specialist NGO’s and CVE practitioners in partnership with Government and or tech sector actors. In such instances, having stock content, archival content, and rapid response creative treatments available at short notice saves both time and resources for already overstretched community partners and stakeholders. Content that can be repurposed and repackaged may stand a better chance of winning the narrative war that ensues in the aftermath of such incidents. The different stages and needs of responding to crisis points also require consideration and awareness in order to ensure consistency and efficacy are maintained.

What Do We Want Our Target Audience to Do Once the Campaign Is Over?

As mentioned previously, a primary aim of community-centric strategic communication is to be able to influence audiences based on a desired end-state or goal. In the CVE context, this focus can take the shape of increased resilience building, rejection of extremism, or greater levels of critical thinking to name a few. When online interventions are allied to relevant offline interventions, the question of longer-term impact or legacy takes on additional significance. Taking the example of having adequate support or redirecting the energy of participants for offline interventions, this allows audiences to seek and access support services or community-linked pathways to social, employment, and cultural inputs that can build a more sustainable impact through longer-term engagement. Within this model of intervention, a key element of the relative success of such efforts is how projects have infused objectives within a community-oriented environment through relevant actors and organizations who are able to engage and retain more “hard-to-reach” groups. Before this model is discussed in the context of the online space, some elucidation of the premise of positive interventions and their proposed outcomes is needed.

A key factor in this discussion is based on the scope, limitations, and bandwidth of communication as

41 Jonathan Ervine, “Five Years on from the Charlie Hebdo Attack,” The Conversation., January 6, 2020, <https://theconversation.com/five-years-on-from-the-charlie-hebdo-attack-je-suis-charlie-rings-hollow-129151>.

42 Mukul Devichand, “How the World was Changed by the Slogan Je Suis Charlie,” BBC News, January 3, 2016, <https://www.bbc.co.uk/news/blogs-trending-35108339>.

43 Joe Parkinson and Drew Hinshaw, “How the ‘Bring Back Our Girls’ Tweets Changed a War in Nigeria,” Wall Street Journal, February 20, 2021, <https://www.wsj.com/articles/how-the-bring-back-our-girls-tweets-changed-a-war-in-nigeria-11613797261>.

a tool to reduce the allure of extremism, harmful socio-political movements, and violence as a means to seek change. While it is acknowledged that radicalization and deradicalization are still very much contested concepts, both serve as useful narratives in their own right to explore both the potential and pitfalls of solely communication-focused interventions. The reason for this suitability is that many CVE communications efforts are based at the conceptual and design level on the propositions advanced by radicalization studies. In order to therefore delve into the longer-term ambitions of such efforts, in terms of impact on audiences and conceptual and real-world variances, anomalies and assumptions need to be identified.

A simple example of the over-reliance on radicalization discourse can be understood when one considers that the concept of socialization is barely mentioned in the literature in regard to possible processes by which to safeguard vulnerable communities from extremism.⁴⁴ Equally, there is an enormous body of literature available that links the allure of extremism, religious cults, violent social movements, and insurrectionist networks to broader themes around the need for identity, a sense of belonging, and a sense of purpose, loyalty, or duty among many who regrettably join such entities. This suggests that if radicalization is seeking to explain how this process occurs, no standard blueprint can be applied as no two journeys into extremism can ever be completely identical. What happens prior to the onset of radicalization is rarely considered, yet in many ways this pre-extremism phase is exactly where many communications efforts in CVE are often basing their objectives (albeit indirectly), as will be explained below.

A UNDP-commissioned study on drivers, incentives, and the tipping point for recruitment by violent extremists in Africa found several common “push” and “pull” factors around locality, opportunity, and forced/exploited incentives at the heart of VEO recruitment.⁴⁵ Such findings are important and need to be understood in relation to the role of social/religious cults and psychology/identity needs also.⁴⁶ The use of identity and belonging narratives need to be better understood in order to inoculate communities against extremist renderings through redirect methods and more open discussion of these themes of these issues.⁴⁷

Whenever a communication-focused intervention is seeking to reduce support for extremist narratives, reduce extremist recruitment and provide alternatives, there are often narratives designed and disseminated that seek to increase knowledge, shape attitudes, and ultimately change behavior among audiences. The main aim is to get audiences to act differently enough to be able to measure some form of reduction in extremist activity, presence, or support. If it is considered that all forms of extremism take root in and around “mainstream” social realities and many adherents of extremist thinking have already made an active choice (voluntarily or otherwise) to reject or move away

44 Paul Hedges, “Radicalization: Examining a Concept, its Use and Abuse,” International Centre for Political Violence and Terrorism Research 9, no. 10 (October 10, 2017): 12–18, <https://www.jstor.org/stable/26351560?seq=1>.

45 “Journey to Extremism in Africa: Drivers, Incentives and the Tipping Point for Recruitment,” United Nations Development Programme, 2017, <https://journey-to-extremism.undp.org>.

46 Daniel Verana, “A Look Behind the Curtain of Cult Psychology,” The State News, October 29, 2020, <https://statenews.com/article/2020/10/a-look-behind-the-curtain-of-cult-psychology>.

47 “Identity, Belonging and Extremism,” European Union: Organising Intercultural and Interreligious Activities: A Toolkit For Local Authorities, n.d., <http://www.congress-intercultural.eu/en/initiative/215-identity--belonging-and-extremism.html>.

from such norms and values, then attempts to dissuade them are offering the same set of norms and values that have been shunned already. In other words, when positive interventions or strategic communication efforts use terminology such as increasing critical thinking, rejecting extremist narratives, and becoming more aware or better active citizens in a strategic objective setting context, they are making a set of assumptions that a narrative on its own can impact and replace the allure of identity, belonging, and sense of duty that extremist ideology offers. This point has two important real-world implications for interventions in today's crowded information space: measurability and specificity.

Extremism offers not only identity, belonging, and sense of duty narratives, but in many instances can offer these through networks, roles, and opportunities for self-gratification. This "package" ultimately replaces traditional notions of family, friendship, and socio-economic stability that form the basis of modern liberal democratic societies, which is in and of itself an issue given that such societies are not the places where the most urgent interventions are needed. Yet for reasons that are beyond the scope of this report to explore in any great detail, this underlying contrast of worldviews still affects how effective positive interventions can be. This point of divergence in understanding and context is then amplified still further when it is considered that the presence of extremist thinking has permeated several mainstream narratives and latched itself on to the rise of disinformation and an increase in harmful actors who are difficult to categorize either as extremists or as activists. Put another way, it has become difficult to target vulnerable groups when mainstream society is now the most vulnerable group and geography plays less of a factor in where needs are most pressing intervention-wise. The loaded nature of the way in which the term extremism is used is in and of itself problematic when seeking to define parameters of interventions, apply legal context, and adhere to human rights.

The rise of technology and the role it plays in offering tools and means by which to proliferate extremist ideas and disinformation are issues that represent another challenge for civil society actors and the tech sector alike, alongside trying to challenge and undermine extremism globally. Conflict zones, areas where autocratic regimes operate, etc., are no longer the sole target for extremists. Between 2016 and 2022 alone, the landscape of influence, partisanship, polarization, and binary narratives has coincided with a sharp rise in new social movements, othering narratives, and cancel culture in many well-developed liberal democracies. How effective can a communications campaign be in this context when other actors and entities offer not just narratives but lifestyles and active participation?

This creates a set of 'relative uncertainties' about some of the intended aims of positive interventions that have already been mentioned in this section. How critically do we want audiences to be able to think? What if they think so critically that they oppose both the narrative being proposed to them and furthermore take steps to voice this opposition? What does resilience look like to someone already being offered a job and a sense of unity and loyalty from a VEO when the social reality they live clearly cannot satiate these needs through "hope" alone? How will having a greater sense of awareness of things negate extremism when levels of trust in authority and questions around legitimacy are now commonplace in state versus citizen narratives? What does citizenship look like in the absence of a clear and identifiable sense of security? What values systems does one draw upon

in considering participation in a VEO?

Such questions raise again the need for positive interventions to be immersed in the lived reality of their intended audiences. Such interventions should rely on highly developed levels of insight and expertise regarding the drivers and factors creating the problem set and able to pursue a suite of nuanced, timely, and compelling narratives that take advantage of the tools on offer within the tech sector (through improved private-public partnerships) in order to offer choice, a sense of digital “sanctity,” and a policy of being present and “there” for audiences in their time of need. Interventions should look to legacy impact as much as initial impact, with as much future-proofing for content and access as resources and scale allow.

The quest for more private-public partnerships within the CVE field again goes to the heart of this debate and is something critical to the long-term success of efforts to counter extremism at the source.⁴⁸ How the role and remit of the CSO sector develop in CVE is seemingly tied to the ways the partnership grows over time to meet the evolving threats, risks, and needs of communities.⁴⁹ YouTube’s Creators For Change program is a good example of how to bring tech and communities closer together in this regard.⁵⁰ Issues linked to the need for trust and credibility are of course paramount to this type of partnership and cannot be overlooked for ‘quick wins’ or vanity level metrics.⁵¹

Section 3: Turning Passive Counter-narratives into Active Strategic Communications

As previously discussed, increased critical thinking as an objective still remains popular within positive intervention design and content threads. The reason for this is based on the proposition that the higher the level of emotional intelligence and ability to think critically about information, choice, and outcomes will offer audiences a safety net from narratives that possess a more binary “them and us” focus. The PeaceGeeks initiative involved critical thinking through elements that include looking, assessing, and responding, then complemented it with Social and Emotional Learning (SEL) to work on emotional intelligence. The rationale is that critical thinking on its own is not enough to foster emotional intelligence.

The generic critical thinking approach champions the role of informed choice making and validating sources as a means to avoid falling into the echo-chamber trap and confirmation bias patterns so prevalent in extremist narrative. In order to make critical thinking a powerful tool in this regard, static

48 Alejandro Beutel and Peter Weinberger, “Public-Private Partnerships to Counter Violent Extremism: Field Principles for Action,” Final Report to the U.S. Department of State (College Park, MD: START, 2016), https://www.start.umd.edu/pubs/START_State_PublicPrivatePartnershipstoCounter-ViolentExtremismFieldPrinciplesforAction_June2016.pdf.

49 “The Role of Civil Society in Preventing and Countering Violent Extremism and Radicalization that Lead to Terrorism,” Organization for Security and Co-operation in Europe, August, 2018, https://www.osce.org/files/f/documents/2/2/400241_1.pdf.

50 “YouTube Creators for Change,” YouTube, September 19, 2016, <https://www.youtube.com/channel/UCYJJpu7FLQqu788cusj6nlg/about>.

51 Kurt Braddock and John Morrison, “Cultivating Trust and Perceptions of Source Credibility in Online Counternarratives Intended to Reduce Support for Terrorism,” *Studies in Conflict and Terrorism* 43, no. 6 (2020): 468–492, <https://www.tandfonline.com/doi/abs/10.1080/1057610X.2018.1452728>.

content and narrative with a one-way engagement focus face an uphill task to permeate both curiosity levels and cognitive avenues to behavior change, simply due to one important component being missing: active two-way engagement. However, this is not the two-way engagement of classical approaches to strategic communication; this is a term used to describe a more immersive, process driven, and decision-making oriented set of exchanges. The main aim of this approach would not solely be the exchange or engagement, but also the experience (ideally a shared experience wherever possible). Ultimately, the human need for shared meaning and experience drives both extremism's allure (at least initially in many cases) but is also the key to challenging harmful narratives through audiences having the chance to immerse themselves in alternative ways of thinking as opposed to just being told about them.

Friction As Counter-speech

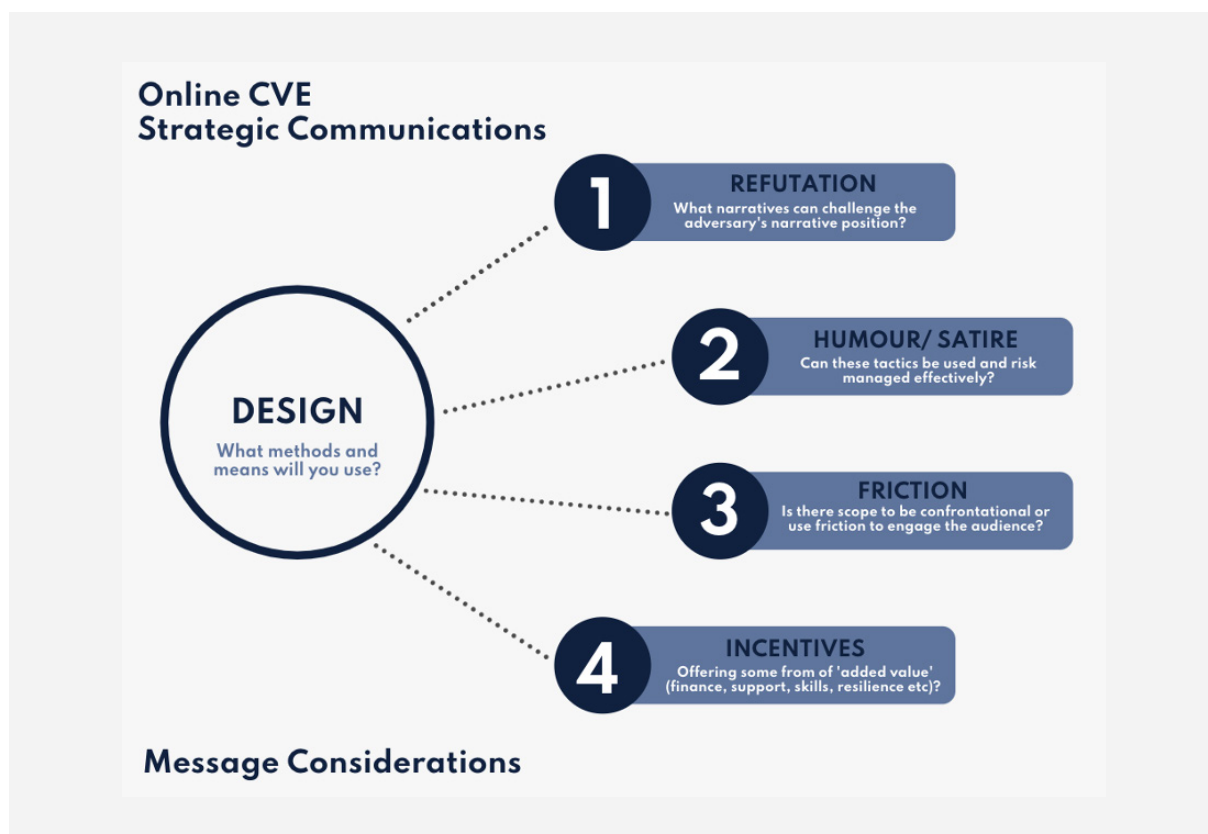
Within the positive interventions space, friction can be used in many ways for different means and end goals (e.g., undermining, refuting, questioning). One aim of using friction in this context is to create uncertainty and doubt about the authority and superiority of the extremist position through direct challenges to the auspices under which such claims are made. This aim is driven by the presence of counter-narratives that question, attack, and malign the assumptions, distortions, and misinterpretations that extremists use, turning these against them through intellectual, rational, yet relatable narrative content that causes “breakpoints” between idea and action in the target audience. The aim is to posit some ideation, but this alone cannot create a total rejection of an entire existing position that may not have used “logic” or “reason” to embed itself originally.

In “Five Considerations for a Muslim on Syria” by Abdullah-X, there are specific trigger points (e.g., questioning of motives, managing emotion, feeding curiosity) used in the narrative that contains elements of friction designed to elicit internal questioning of the justifications presented by ISIS as well as broadening out the conflict with issues related to a desire for power and control.⁵² The goal of such a narrative is to give the user an opportunity to question his/her motives for considering harmful actions or ideas alongside being given intellectual tools to be able to better understand the “bigger picture.” According to Kurt Braddock, “Messages like the one communicated in this video provide viewers with a source of pride that (a) they can identify with and (b) challenge terrorist ideologies and actions... Abdullah-X, who describes himself as having the ‘mind of a scholar’ and the ‘heart of a warrior’, shows young Muslims targeted by ISIS propaganda that they can be proud of their Muslim heritage, but need not affiliate with ISIS.”⁵³

Various tactics within this context can be used in isolation or as a set of content within the same project/campaign context, as demonstrated by the figure 5 below:

.....
52 Abdullah-X, “Five Considerations for a Muslim on Syria,” Youtube.com, March 7, 2014, Ibid, <https://youtu.be/tKKbydB4scA>

53 Kurt Braddock, *Weaponized Words: The Strategic Role of Persuasion in Violent Radicalization and Counter-Radicalization* (Cambridge, UK: Cambridge University Press, 2020), 196.



Another use of friction in counter-speech is to “call out” extremists by isolating their tactics and actions as those of people who clearly do not value human life and the values that let societies flourish as opposed to perish. This works best when the positive examples are from within the same social, cultural, historical, religious, or geographical context. Such tactics can utilize “discrete emotion theory” to communicate messages that offer alternative yet compelling pathways to extremism through positive goal targeting, successful actions, and highlighting similarities between these approaches and the target audiences’ core values. These examples suggest that friction can have subtle variances in how it is used, applied, and interpreted. One way to identify this tactic is by labeling such messaging as “positive friction,” an example of which can be found in the “Syria Street Stories” campaign.⁵⁴

Positive friction is perhaps best illustrated in the commercial advertising world, where the constant struggle between sportswear giants Nike and Adidas for market dominance uses differentiation and lifestyle narratives to offer “unique” choices to often similar customer bases.⁵⁵ Companies “challenge” their rivals through unique storytelling and values-based narratives, all the while undermining the same things in their rivals by identifying areas of divergence in image and lifestyle, etc. Nike has long maintained that “just do it” is about being both in the moment and an unapologetic mindset, in order to harness, access, and develop one’s capability levels. Adidas, which also targets a similar customer profile, like to remind these potential customers about their brand’s longevity, street cred, and cutting-edge persona (in terms of being both ‘old-school’ and innovative), using influencers and familiar faces to build brand value. They have a well-known strap-line “the brand with the three stripes,” despite

54 “Abbas’s Story,” YouTube.com, February 26, 2016, <https://www.youtube.com/watch?v=4WeUQy972Kw&list=PLb3EHFgagz0xiR2tRNq-qahBdEH-hW49-s>.

55 Sofia Deneke, “Nike vs. Adidas: Whose Marketing Strategy Reigns Supreme?,” Trig, July 27, 2017, <https://www.trig.com/tangents/nike-vs-adidas>.

their actual logo having changed from the classic logo the strap-line is referring to. Both brands have strap-lines that mean different things, but still speak of “values” that most likely mean similar things in both the lifestyle and sports contexts. The friction element is hidden inside sophisticated and often indirect references to each other’s brand without directly naming the other party.

What Is an Appropriate Amount of Friction?

This relates to both the appetite for the medium and the propensity to manage and mitigate risk. Direct confrontation through friction is technically designed to elicit a response, and this means that actors need to gauge and model the potential impact of friction on reputation, safety, and audience wellbeing levels. The rationale at use here is that friction begets more friction and this means having a willingness to be reactive, flexible, and attentive to changing dynamics and subtle cues. This is a scenario that must include comprehensive end-state planning and what decisive conditions need to be met in order for the use of friction to be considered a successful tactic and not a “flop,” risk, or liability.

Therefore the use of friction carries both specific potential benefits while also significant risks. From a management perspective, friction runs the risk of snowballing and taking on a life and narrative angle of its own. If this includes the target audience becoming embroiled in a war of narratives and attacks on the other party, online safety issues quickly translate into offline risk of harm issues, given online identity is not always as hidden from prying eyes as people like to think. Project administrators in this case may need to have in place support structures and contact information for audiences to use in case of threats and trolling/harassment, as well as encouraging more digitally “savvy” account names and profiles, etc.

Managing and Measuring Impact of Friction on Audiences

In terms of measurement and impact, A-B testing with and without elements of friction in the content on the same audience offers an empirical way to demonstrate both the reach, engagement, and second level or sub-effects of the friction being employed. In some more sensitive operating environments, multiple channels carrying subtle variations in messaging and tone can also create a network effect for friction-based content that suggests support for the antagonist’s point of view is smaller than support for the protagonist’s perspective. Pooling the bandwidth of such content in terms of data mining for where the friction-based content appears in different online places also allows for reach and impact to be measured more widely than just on the official channels.

Building User Resilience in the Online Space

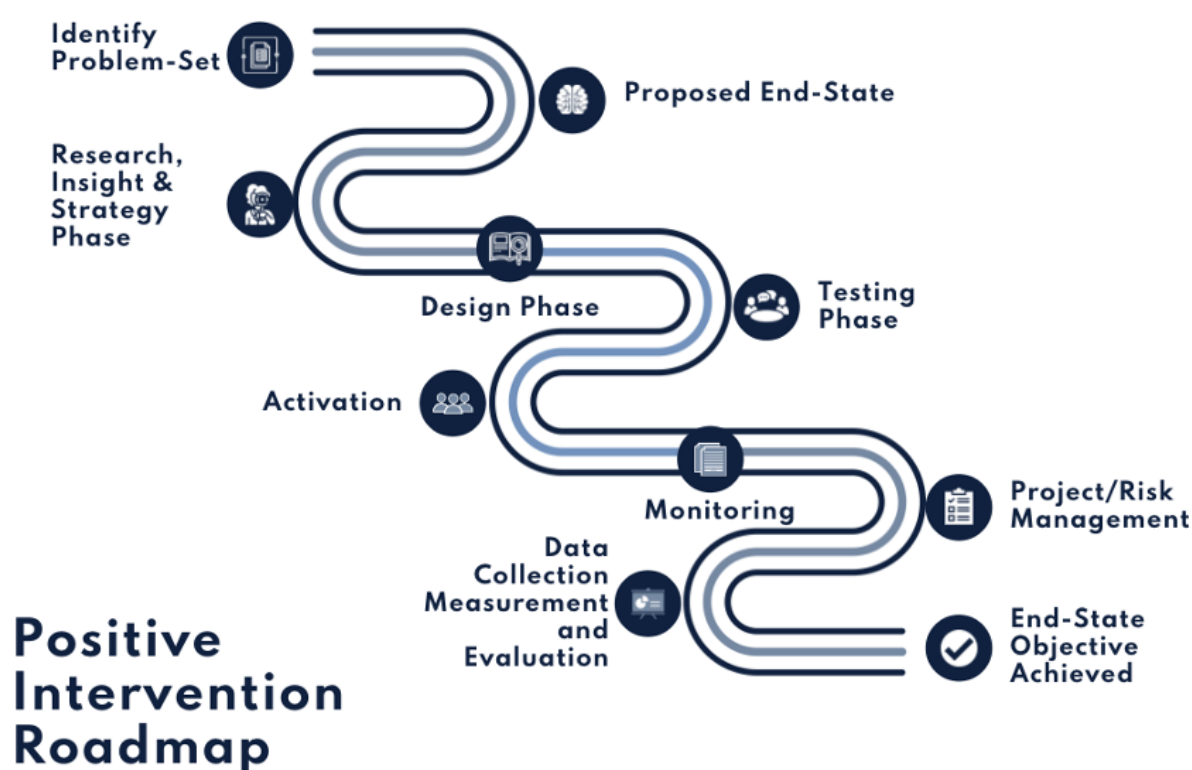
The presence of the term resilience building in the online intervention space becomes more of a long-term success criterion or factor when it is part of a drive to empower audiences to adopt peer support, mentoring, and self-reliance narratives alongside having prebunking, counter, or alternative examples to lean on. These processes not only have positive effects on engagement and trust but also build self-protection and community self-regulation and management into the mindset of audiences who may be very keen on doing more and being part of something bigger. This approach

directly fosters both resilience and empowerment alongside offering tangible measurement opportunities for evaluation purposes.

As an obvious contrast, resilience in the offline context is something that requires building, constructing, and layering support, capacity building, resources, partnerships, and infrastructure over a sustained period of time. As such, it is good practice to view this approach as distinct to direct counter-narrative or counter-extremism efforts, where there is an identified need to address deep-rooted ideological issues manifested within communities. This is where the role of aspiration as a form of incentivization within online resilience-building efforts can be useful to keep audiences engaged over time.

The tactics, considerations, and challenges of deploying these various methods highlight the need for positive interventions to be based on process and insight-driven stages, phases, and pathways. The more in-depth the insight and planning processes are, the greater the likelihood of the creative and content dissemination phases will yield positive results for the overall project's objectives/end-state needs. To illustrate this point, the roadmap below amalgamates the processes and phases that are key components of an effective positive intervention:

Figure 6: Positive Interventions Process



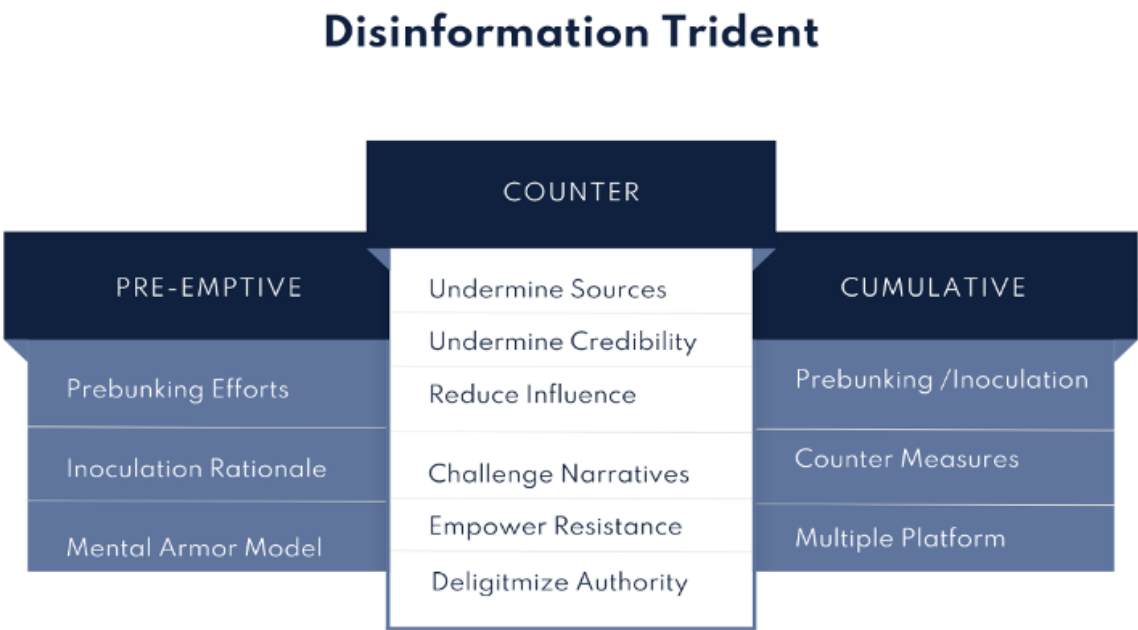
What Can We Learn from the Disinformation Space?

Extremist narratives and disinformation are surprisingly similar from the point of view of having a

symbiotic relationship. It can be argued that extremism is a branch and disinformation its root, in that both need each other to thrive and evolve. In terms of the online space, different emergent and historical factors come into focus regarding the role of tech and civil society that require discussion and contextual clarity in terms of learning, best practices, and pitfalls. An obvious place to start is with the assertion that disinformation is not new, and like extremism has morphed to mean different things at different times in history. What is new is that the proliferation, scale, and size of disinformation are at unprecedented levels and permeate all aspects of social and digital life.

Online disinformation is not limited to one subject matter or context. Advances in technology and the use of tech tools have made huge inroads into tackling disinformation, misinformation, and propaganda through data mining, machine learning, and advanced algorithmic capabilities. The issues remain those of scalability and volume of response against scale and dissemination of disinformation. Experts maintain that information is the new weapon of choice for harmful actors and VEOs. Terminology has appeared in recent times that suggests this issue is taking up more and more thinking space for government, tech sector, and civil society. Terms like “narrative warfare,” “durable disorder” and “weaponized words” are just some examples of how ubiquitous the issue of disinformation has become.⁵⁶ What can the CVE sector learn from approaches to tackling disinformation and other forms of harmful information online? Essentially, this battle for information influence and the weaponizing of narratives has prompted tech-centric actors to initiate the use of tools to identify, target, take down, refute, and challenge disinformation through a more direct relationship with end-users. What this translates into looks like the following trident:

Figure 7: Disinformation Trident



.....
 56 Caroline Jack, “Lexicon of Lies: Terms for Problematic Information,” Data Society, August 9, 2017, <https://datasociety.net/library/lexicon-of-lies/>.

Pre-emptive counter-disinformation involves the use of take downs, censorship, and removal from platforms (account suspension). Active counter-disinformation involves using algorithms and machine learning to warn and engage end-users when they are interacting with disinformation, fake news, or alike, creating a choice dilemma for the end-user and offering the chance to learn more about the threat posed by disinformation. Cumulative counter-disinformation is based on taking fact-checking, critical thinking, rational choice modeling techniques, and offering alternative sources of information by redirecting end-users to other places by treating them with an empathetic manner to build trust and maintain it.

A key lesson from the disinformation space is that it is always morphing and oscillating on different wavelengths in order to attract and then retain attention. Disinformation plays a “long game,” basing itself on a strategic plateau that seeks to build up momentum over time in order to undermine rational thinking, objectivity, and trust in traditional modes of authority with one overall end-state in mind – the person becoming the disinformation. These different wavelengths refer to the use of different means, mediums, and technologies to confer similar meanings but with subtle variations. Examples include memes, humor, satire, deepfakes, animations, trolling, and friction. These may all have different visual facets but carry the same underlying narrative on whatever the subject matter is. The aim here is to attract users with at least one of these “hooks” in order to then develop a longer-term relationship to a point where this user is now actively creating, sharing, and adhering to the desired narrative.⁵⁷ The use of “alternative facts,” canceling, and trolling techniques using bots start to create a chasm in the thinking space of the end-user, which can often lead to cognitive dissonance. Another learning from this context is that disinformation is very much a mind-game pursuit in its initial stages, as it seeks to replace rational thinking and emotional intelligence with conspiratorial ideas and groupthink.⁵⁸ Extremism in some guises also adopts a similar tactic through narratives of guilt, emotional cues, sentiment generation, and redemption arcs, eventually creating a “them and us” mindset. Some relevant case-study examples that address the nexus between disinformation and extremism can be found in the work by Moonshot CVE (see below).

Two recent projects aim to counter disinformation, political polarization, and violent extremism by encouraging internet users to think more critically about the information they consume. These campaigns prompted internet users to consider how their behavior affected their online peers and offered detailed resources or one-to-one support on how to hold constructive conversations that bridge ideological and political divides. This section provides an overview of these projects and how they engaged online audiences beyond delivering a counter message or alternative content.

Using Gamification to Advance Media Literacy in Indonesia

Moonshot has monitored the disinformation environment in Indonesia since 2019 and developed a database of key disinformation narratives and the audiences consuming them. In partnership with the University of Notre Dame, the International Research & Exchanges Board (IREX), and GeoPoll, they

.....
 57 “Prebunking Anti-Vaccine Narratives: An Effective Alternative to Debunking Individual False Claims,” Jigsaw, March 2, 2022, <https://medium.com/jigsaw/prebunking-anti-vaccine-narratives-an-effective-alternative-to-debunking-individual-false-claims-78f0047a8b47>.

58 “Hate Clusters Spread Disinformation Across Social Media. Mapping Their Networks Could Disrupt Their Reach,” Jigsaw, July 28, 2018, <https://medium.com/jigsaw/hate-clusters-spread-disinformation-across-social-media-995196515ca5>.

used this research to deploy digital campaigns and reach vulnerable audiences with a media literacy website and gamified content.

The website, Literata.id, is designed to guide new digital arrivals who encounter disinformation on the internet and improve their media literacy skills. Literata.id contains eight lessons from the IREX curriculum, including how to spot echo chambers, how the language used in news articles can mislead readers, and how to identify different types of disinformation. Each lesson includes a video and a short quiz to test users' engagement and understanding of the content.

Figure 8: The Literata.id website



Between September 2020 and April 2021, Moonshot developed and tested a media literacy game called Gali Fakta designed for the same audience. The script transposed lessons from the IREX curriculum and was built around real examples of disinformation from Indonesia.

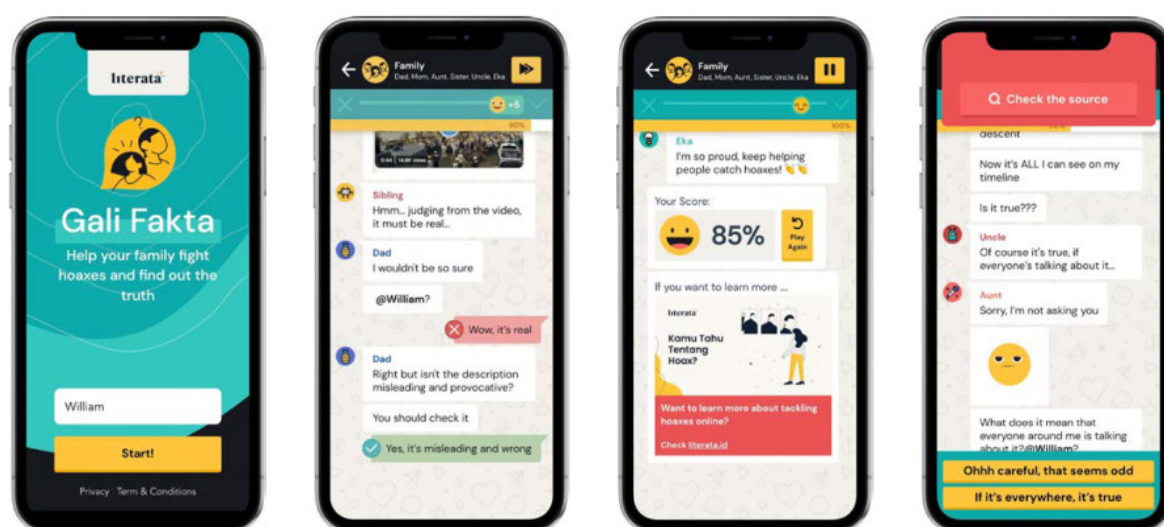
The decision to use real disinformation in the game was informed by the inoculation method developed by researchers at the University of Cambridge. The method is based on the theory that psychological resistance to disinformation can be developed by exposing individuals to weakened versions of fake or manipulated stories that they will come across in the real world. Moonshot exposed users who had previously engaged with disinformation with select examples in prepared and carefully managed settings. All examples included a disinformation warning or a rapid feedback loop so that players were clear on what was true and what was false. Moonshot also created a second version of the game with low-risk examples of disinformation for the purposes of general consumption.

How the game worked for users

Gali Fakta is designed in the style of a family WhatsApp chat. This familiar and entertaining scenario reflects the real-world context of how disinformation spreads in Indonesia, where WhatsApp is the most popular messaging application. Family chat groups make up over 70% of Indonesian user activity on the platform. A similar approach can be seen in MediaSmarts work focused on youth-focused programming.⁵⁹

Figure 9: Gali Fakta Inoculation Game

Gali Fakta: The disinformation inoculation game



The game teaches IREX's media literacy lessons through leading questions and social proof. When a user is prompted to spot disinformation, correct answers are rewarded by points or the family reaching a consensus. Incorrect answers are docked points and met with general confusion by family members. The player's "cousin," Eka, is a media literacy expert and functions as a corrective voice should the player get any answers wrong. In addition to Eka stepping in to correct them, players who answer incorrectly are also immediately given the chance to correct themselves. Should they stick to their original incorrect answer, they lose more points and receive an explanation from Eka.

Between April and October 2021, Moonshot ran campaigns on Twitter, Google Search, and Google Display to reach audiences in Indonesia. They simultaneously advertised the website and game, and when these campaigns were complete, Moonshot used post-surveys to measure whether engagement with the media literacy content had impacted participants' self-reported intentions to respond responsibly and proactively to online disinformation.

59 "Reality Check: The Game," Media Smarts: Canada's Centre for Digital and Media Literacy, n.d., <https://mediasmarts.ca/digital-media-literacy/educational-games/reality-check-game>.

How users interacted with the media literacy content

Overall, 24,581 individuals visited the website, of whom 72 took a quiz and 289 watched a video. 8,128 individuals viewed the game page, of whom 781 started playing it and 98 completed it. The website had more visitors and a lower bounce rate than the game. On the website, users could access content right away, whereas the game requires the user to type in a username and commit “10 minutes of their time.”

Despite the website's lower bounce rate, the game was more effective at engaging users and maintaining their attention. Game players spent approximately 12 times (1,185%) as long engaging with the media literacy content than those who visited the website without playing the game. Results were statistically significant at the 95% confidence level ($p < .05$).

However, the impact of media literacy content on behavior change was inconclusive. Differences between control and treatment groups taking the survey made it impossible to measure whether or not there was a statistically significant behavior change, as the control group members self-reported as being significantly younger and more highly educated. When these demographics were controlled for, the sample size was too small to conclude if the game increased players' media literacy.

How this approach can be improved

This outcome raises a number of considerations for future programming. A key lesson is the need to more robustly connect user behavior to survey responses. Moonshot's design did not connect data on user behavior from Google Analytics with the users who responded to surveys. Attributing multiple behaviors to a persistent user in this way is challenging, but the stronger the link between user behavior and the results of their post-treatment survey, the more likely a program is to confidently measure any changes in behavior.

The program has been extended for four years to enable further adaptation and testing. Moonshot is exploring a number of possible solutions with its partners, including new survey software which would enable pseudonymous connections between user behavior and pre and post-survey results, and collaborating with GeoPoll to SMS or call users after their treatment.

Preventing U.S. Election Violence Through Strategic Messaging

In November 2020, Moonshot launched a new model for election violence prevention and peacebuilding. Over four months, the campaign sought to counter and reduce the threat of violence in the aftermath of the 2020 U.S. presidential election by safeguarding internet users seeking or consuming content that could incite violence and offering alternative pathways that supported shared values, inclusive citizenship, and mental health.

How the campaign was designed

Key audiences included individuals engaging with QAnon, white nationalism, violent armed groups

such as the Proud Boys and Oath Keepers, and disinformation related to the results of the 2020 presidential election. The campaigns drew more than 1.7 million engagements from at-risk internet users, including retweets, clicks, video views, and downloads.

With input from local community partners, Moonshot designed and tested 21 pieces of original content containing de-escalation messages and promoting shared values, and redirected thousands of individuals to crisis counseling services. The campaigns' messaging, content, and targeting were refreshed continuously, based on an open-source analysis of post-election threats and partner consultations. Moonshot also surged de-escalation advertising at critical moments, such as the January 6 attack on the Capitol.

Moonshot's campaigns ran on four advertising networks: Google's Search, Display, and YouTube platforms, as well as Twitter. Beyond measuring standard indicators of audience engagement, such as ad impressions, views, and clicks, Moonshot evaluated individuals' engagement with partner websites, providing mental health support, guides on having difficult conversations with peers and loved ones, and promoting active listening skills.

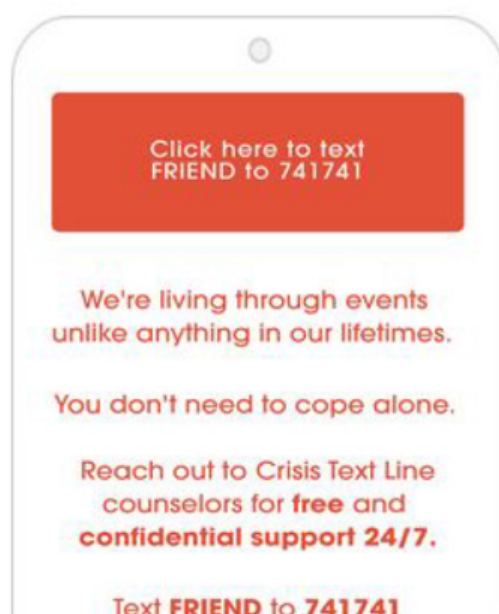
Impact and reach

Over 22,000 users engaged with an ad offering psychosocial support and visited websites featuring mental health resources or the Crisis Text Line (CTL) service. Thirty-three individuals texted the CTL helpline using a unique referral keyword, resulting in a total of 39 conversations, and it is likely that a greater number reached out using the generic keyword on the service's primary website (see Figure 10). This outcome provided evidence that connecting at-risk individuals with mental health support services can result in sustained engagement. To date, 163 conversations have occurred through Moonshot's partnership with CTL.

5,500 users visited a website focused on election integrity. The website was advertised via messaging about the importance of having healthy civic conversations with Americans holding opposing points of view. Over 100 visitors downloaded and shared bespoke discussion guides with titles like "How to Talk to Someone You Disagree With" and "How to Talk about the Election and Move Forward." Visitors downloaded 130 PDF guides in total. This provided evidence that Moonshot's audiences actively viewed and saved resources that fostered positive civic engagement and challenged polarization and hate. Audiences interested in election disinformation were most likely to visit the website and download the most content.

1,250 users visited a website featuring actionable steps on how to process difficult emotions and build a supportive local community. This website was advertised via messages about the importance of neighborliness and inclusive American identity. 18 individuals interacted with sections of a "Get Involved" page, which provided information on how to process and express emotions in a healthy way, how to care for themselves and other individuals, and how to solve challenges collectively. Visitors also reviewed a "conversation prompts" page, providing information on how to have productive conversations with other Americans.

Figure 10: Crisis Line Microsite. Visitors who clicked the red button were instantly redirected to their preferred messaging application to start a conversation with a counselor.



Overall, Moonshot found that offering psychosocial support is a uniquely effective way to engage audiences at risk of participating in political violence. While messages were at times met with skepticism and occasional hostility, more often content was engaged with positively, and shared and endorsed enthusiastically. Through the partnership with CTL, Moonshot and its partners were able to facilitate conversations with crisis counselors for 33 Americans, suggesting that members of these audiences are indeed open to repeated engagement and may continue to use psychosocial resources after initial exposure. This also indicated the value of involving mental health organizations in future projects.

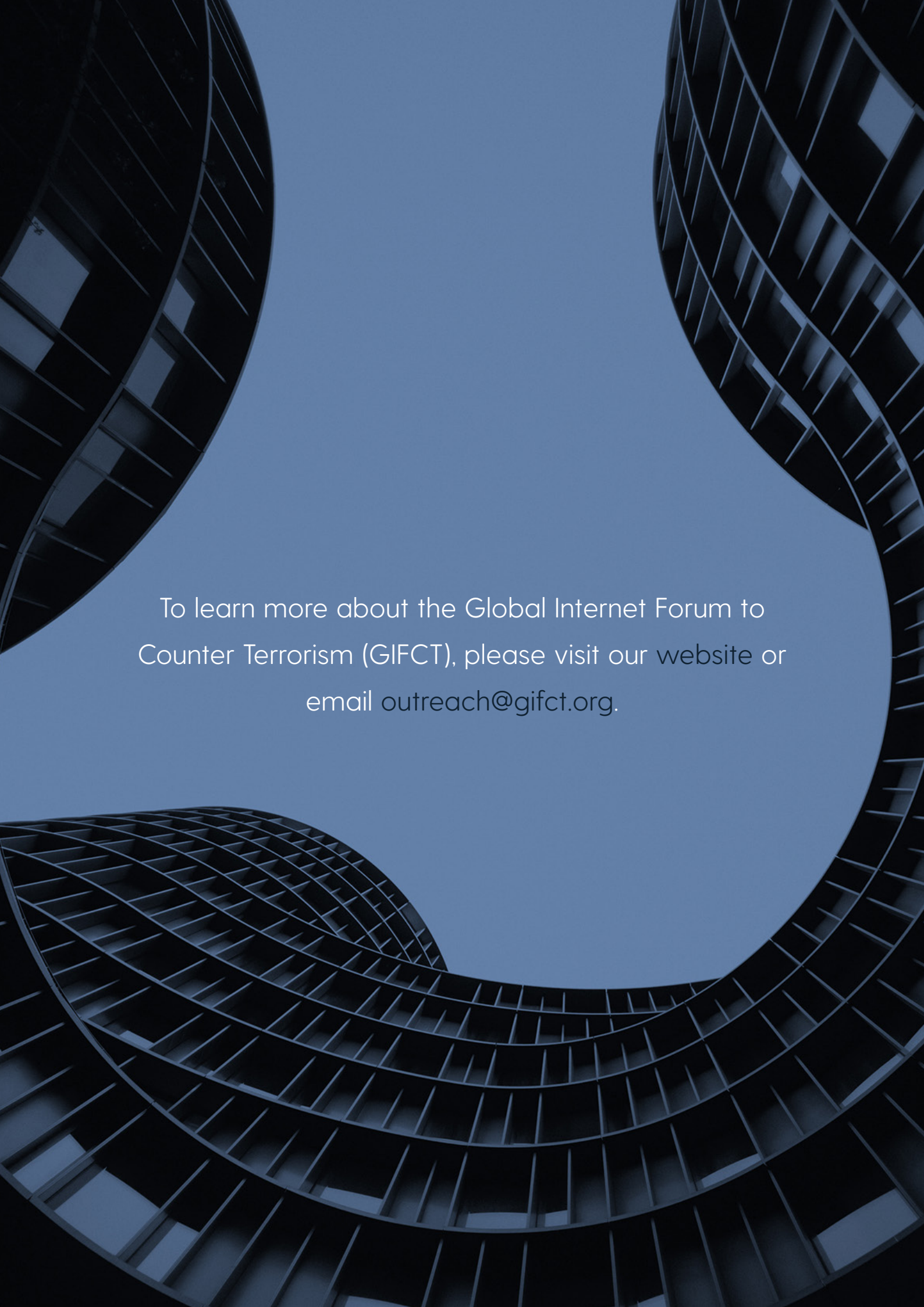
The difficulty in directly applying learning from the disinformation space to the CVE space is that where race, religion, and creed/doctrine-based narratives are present, the skill in using tech tools, creating warnings, and building cumulative resistance to harmful ideas becomes fraught with confusion over representation, authority, and credibility. However, as findings from pilot testing in prebunking approaches begin to emerge and help to shape future plans, machine learning capability continues to gather at pace, and a greater effort is made to align the private-tech sector with the CSO sector, is it reasonable to expect a much larger footprint of pre-emptive, active and cumulative actions to become commonplace within online CVE efforts. The age of information warfare brings with it the possibility that merely winning the narrative battle to be first or "own the narrative" will not be enough, because without discrediting and fully undermining the opposing narrative, ideas endure simply because they can now be repurposed to become relevant again and more durable. The emerging set of considerations seems to be how to work at a mass scale in the pre-emptive space while simultaneously ensuring the micro-targeting tools are fit for purpose from a moral, ethical, and impact standpoint.

Conclusion

The convergence and continued growth of narratives that espouse binary and otherization facets within their central ideas mean 'polarization' and 'unity' alike are now tools more easily weaponized through communication engagement and lifestyle choices. Whereas CVE has morphed into its own industry with accepted norms and practices, VEOs, terrorist networks, and their supporters have often been able to harness technology and engagement to their advantage, while the policy, tech, and civil society sectors often struggle with how best to use these same tools within the confines of legal, social, and ethical considerations.

This report sets out key learning on central issues linked to effective strategic communication in the online space, and in doing offers both opportunities and challenges to the sector. The presumed rise of disinformation (which has always been present and will always be) takes the focus away from how this tactic is used by extremist groups and instead concentrates too heavily on its ontological roots. Visibility of disinformation is a distinct issue to the use of disinformation, as is the presence of extremism in relation to the proliferation of extremism. The prebunking approach can make huge strides in not only tackling conspiracy narratives but in building more effective prevention inputs for the CVE space. The use of influencers within CVE has become meshed alongside the ongoing debates around using "credible" voices, and this context needs further research and some form of conceptual framework development in order to be better understood.

The strategic and tactical considerations highlighted in this report outline the need for more emphasis to be placed on understanding audience behavior(s) better and at a more granular level. The relationship between problem sets, strategy, and end-state objectives must be more closely aligned to the use of innovation, offline engagement learning, and diversity in tactical approaches (e.g., humor, incentives, and friction) that have the potential to enhance content as well as engagement. These issues relate at the macro-level to the continued need for measurement and impact assessments to be based on a more nuanced understanding of the role sentiment, semiotics, individual eco-systems, and legacy impact on both campaigns and their recipients. The presence of diverse and often hidden harmful actors in the online space means the tools offered by the tech sector are the key to undermining hate, extremism, and violent extremism, but also point to the need for increased private-public partnerships to sit at the heart of CVE efforts. This report has highlighted that a key tool for this sector to effectively counter violent extremism is the relationship basis between those skilled in the practice of CVE and those equipped with the resources to help deliver results.



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