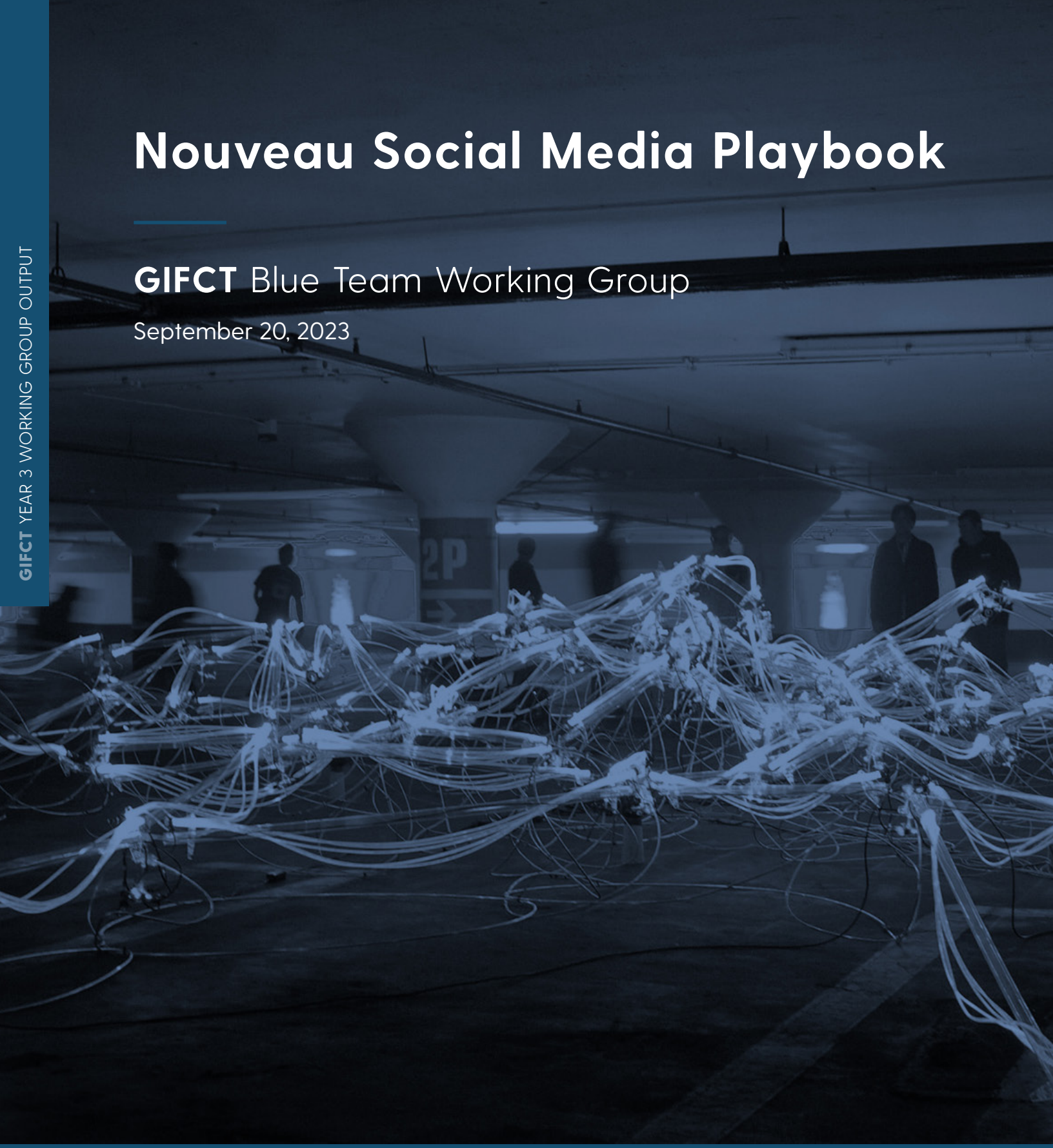
GIFCT YEAR 3 WORKING GROUP OUTPUT

Nouveau Social Media Playbook

GIFCT Blue Team Working Group

September 20, 2023



GIFCT
Global Internet Forum to Counter Terrorism

Dr. William Allchorn
Richmond, the American International University in London

About GIFCT Year 3 Working Group Outputs

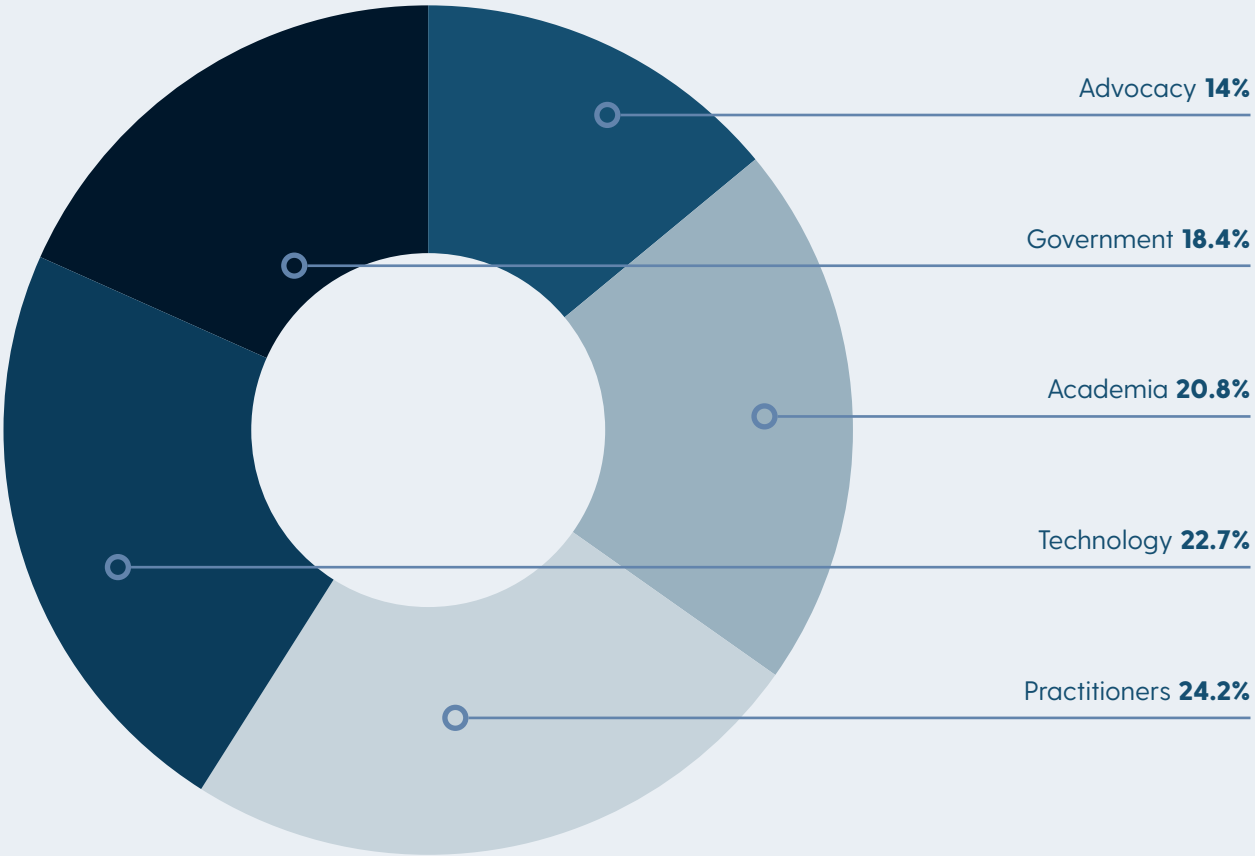
By Dr. Nagham El Karhili, Programming and Partnerships Lead, GIFCT

In November 2022, GIFCT launched its Year 3 Working Groups to facilitate dialogue, foster understanding, and produce outputs to directly support our mission of preventing terrorists and violent extremists from exploiting digital platforms across a range of sectors, geographies, and disciplines. Started in 2020, GIFCT Working Groups contribute to growing our organizational capacity to deliver guidance and solutions to technology companies and practitioners working to counter terrorism and violent extremism.

Overall, this year's five thematic Working Groups convened 207 participants from 43 countries across six continents with 59% drawn from civil society (14% advocacy organizations, 20.8% academia, and 24.2% practitioners), 18.4% representing governments, and 22.7% in tech.

WG Participants

Sectoral Breakdown



Beginning in November 2022, GIFCT Year 3 Working Groups focused on the following themes and outputs:

1. **Refining Incident Response: Building Nuance and Evaluation Frameworks:** This Working Group explored incident response processes and protocols of tech companies and the GIFCT resulting in a handbook. The handbook provides guidance on how to better measure and evaluate incident response around questions of transparency, communication, evaluation metrics, and human rights considerations.
2. **Blue Teaming: Alternative Platforms for Positive Intervention:** After recognizing a gap in the online intervention space, this GIFCT Working Group focused on highlighting alternative platforms through a tailored playbook of approaches to further PVE/CVE efforts on a wider diversity of platforms. This included reviewing intervention tactics for approaching alternative social media platforms, gaming spaces, online marketplaces, and adversarial platforms.
3. **Red Teaming: Assessing Threat and Safety by Design:** Looking at how the tech landscape is evolving in the next two to five years, this GIFCT Working Group worked to identify, and scrutinizes risk mitigation aspects of newer parts of the tech stack through a number of short blog posts, highlighting where safety-by-design efforts should evolve.
4. **Legal Frameworks: Animated Explainers on Definitions of Terrorism and Violent Extremism:** This Working Group tackled questions around definitions of terrorism along with the impact that they have on minority communities through the production of two complementary animated videos. The videos are aimed to support the global counterterrorism and counter violent extremism community in understanding, developing, and considering how they may apply definitions of terrorism and violent extremism.
5. **Frameworks for Meaningful Transparency:** In an effort to further the tech industry's continued commitment to transparency, this Working Group composed a report outlining the current state of play, various perspectives on barriers and risks around transparency reporting. While acknowledging the challenges, the Working Group provided cross sectoral views on what an ideal end state of meaningful transparency would be, along with guidance on ways to reach it.

We at GIFCT are grateful for all of the participants' hard work, time, and energy given to this year's Working Groups and look forward to what our next iteration will bring.

To see how Working Groups have evolved you can access Year One themes and outputs **HERE** and Year Two **HERE**.

Nouveau Social Media Playbook

Dr. William Allchorn[1]

The GIFCT Blue Team Working Group (BTWG) explored how best to facilitate filling a gap in the online intervention space. Practitioners working on preventing and countering violent extremism (PVE/CVE) tend to use only three to four larger social media platforms for most intervention efforts. This BTWG developed this Playbook focused on highlighting alternative platforms for potential positive interventions. The result is a tailored playbook of approaches to further PVE/CVE efforts on a wider diversity of platforms. It aims to help activists in their own efforts to challenge hate and extremism online and foster wider CSO-Tech Company partnerships looking at intervention potentials on (1) nouveau social media platforms, (2) gaming platforms, (3) lifestyle and marketplace platforms, and (4) adversarial platforms. It also includes a chapter on (5) regional and cultural sensitivities for positive interventions. All Working Group outputs are made available on the GIFCT Working Groups page.

Introduction

Over the past decade, several adversarial shifts have occurred in response to deplatforming efforts that have made Terrorist and Violent Extremist (TVE) actor engagement with social media more dispersed, nuanced, and enigmatic.[2] This has occurred at the same time that the internet and social media have largely supplanted both traditional forms of media and face-to-face encounters in extremist efforts to spread messages of hatred, which has enabled sophisticated and targeted propaganda techniques to recruit the right audience.[3] The global radical right is no exception to this trend. For example, as seen through the attacks in Poway, Christchurch, El Paso, Halle, and Hanau,[4] the use of manifestos,

1 Dr. William Allchorn is an Adjunct Associate Professor in Politics and International Relations at Richmond, the American University in London and Honorary Senior Research Fellow at the Policing Institute for the Eastern Region, Anglia Ruskin University. His email is allcho@richmond.ac.uk.

2 See Julia Ebner, "Counter-Creativity: Innovative Ways to Counter Far right Communication Tactics," in Post-Digital Cultures of the Far Right: Online Actions and Offline Consequences in Europe and the US, eds. Maik Fielitz and Nick Thurston (Bielefeld: Transcript Verlag, 2018), www.transcript-verlag.de/en/detail/index/sArticle/4371?number=978-3-8394-4670-6.

3 William Allchorn, "Technology and the Swarm: A Dialogic Turn in Online Far right Activism," GNET, January 17, 2020, https://gnet-research.org/2020/01/17/technology-and-the-swarm-a-dialogic-turn-in-online-far right-activism/.

4 Phil Helsel, "Suspect in Christchurch mosque shootings charged with terrorism," NBC News, May 21, 2019, www.nbcnews.com/news/world/suspect-christchurch-mosque-shootings-charged-terrorism-n1008161; Vanessa Romo, "El Paso Walmart Shooting Suspect Pleads Not Guilty," NPR, October 10, 2019, www.npr.org/2019/10/10/769013051/el-paso-walmart-shooting-suspect-pleads-not-guilty; "German Halle gunman admits far right synagogue attack," BBCI, October 11, 2019, www.bbc.co.uk/news/world-europe-50011898; Sarah Hucal, "Racially motivated terror attack in Hanau puts Germany's right wing extremism into focus," ABC News, February 27, 2020, https://abcnews.go.com/US/racially-motivated-terror-attack-hanau-puts-germanys-wing/story?id=69128298.

online meme culture, and conspiracy theories can lead to powerful offline effects,[5] with the seemingly sporadic and "solo" actor nature of the attacks masking a broader toxic online network of religiously, ethnically, and racially motivated hatred.[6]

One frontier that has largely escaped the attention of preventing and countering violent extremism (P/CVE) scholars and practitioners are more ephemeral audio and short-form video-based features within platforms such as Instagram Stories, TikTok, Clubhouse and Facebook Stories—what we're defining here as 'Nouveau Social Media' (NSM).[7] This playbook attempts to address this lacuna.

Based on (a) interviews with the Meta, TikTok, and Clubhouse platform teams, (b) focus groups that include a large array of P/CVE professionals from the GIFCT Blue Team Working Group, and (c) a survey of existing Nouveau Social Media P/CVE efforts, this NSM playbook helps define recent adversarial shifts and possible effective interventions on these types of surfaces. We will then conclude by proposing possible recommendations of scientifically-proven, sustainable, and scalable P/CVE efforts that could be applied elsewhere.

Section I: Background

Introduction

After key enforcement efforts by mainstream social media platforms against TVE actors over the past eight years, such actors have responded by co-opting less overt and more coded and circuitous platform tactics in order to recruit like-minded individuals and spread their ideological propaganda and hatred. A key inflection point identified across all NSM surfaces identified below is a transition by TVE actors from being less extreme in public-facing spaces while being more extreme in private-facing spaces.[8] Mainstream platforms like Facebook, Instagram, TikTok and Clubhouse have also seen attempts at recruitment where extremists try to co-op so-called 'normie' behaviors or 'normie appearing' accounts in order to circumvent deplatforming and defuse ideological content on mainstream social media—with such "raids" coordinated on encrypted or less-regulated social media before being implemented.

5 For a good overview of the 2019 wave of global far right extremist attacks, see Graham Macklin, "The El Paso Terrorist Attack: The Chain Reaction of Global Right-Wing Terror," CTC Sentinel, December 2019, https://ctc.usma.edu/app/uploads/2019/12/CTC-SENTINEL-112019.pdf.

6 William Allchorn, Moving Beyond Islamist Extremism: Assessing Counter-Narrative Responses to the Global Far Right (Stuttgart: Ibidem/Columbia University Press, 2022).

7 Nouveau Social Media platforms include surfaces (i.e., places where user content takes place that could/might be possible to intervene in, including newsfeeds, direct or group messaging, targeted ads or suggested content, search results, platform education messaging to users) that privilege the use of more short-form video (e.g., TikTok), ephemeral content (e.g., Facebook/Instagram Stories and Reels), and audio chat rooms (e.g., Twitter or Instagram Audio, or Clubs or Houses) that have recently emerged in the social media space.

8 For a further exploration of this in relation to the far right, see William Allchorn, "Beyond Islamophobia? The role of Englishness and English national identity within English Defence League discourse and politics," National Identities 21, no. 5 (2019): 10.1080/14608944.2018.1531840; Matthew Feldman and Paul Jackson, Doublespeak: The Rhetoric of the Far Right Since 1945 (New York: Columbia University Press, 2014); Paul Jackson and Matthew Feldman, The EDL: Britain's 'New Far Right Social Movement (Northampton: Radicalism and New Media Group, 2011).

What Different NSM Surfaces Exist & How TVE Actors Exploit Them

Surface 1: Short-Form Audio & Video (e.g., Facebook/Instagram Stories, Reels & TikToks)

TikTok has become a key destination for short-form mobile video and as a result has also become a key destination for extremist actors of all stripes to experiment with in recent years. For example, on November 26, 2020, a Pakistani imam from a small town north of Paris was sentenced to 18 months in jail and expelled from France for posting videos on TikTok praising recent jihadist attacks that happened in the country and celebrating the attackers.[9] Systematic scans of TikTok content revealed hundreds of postings related to far right extremist ideology (e.g., posts and streams about fascism, racism, antisemitism, anti-immigration, chauvinism, nativism and xenophobia) and glorifying far right lone actor terrorists (e.g., Breivik, Tarrant, Roof, and Rodgers).[10] Early forays by TVE actors on audio platforms have included so-called 'beta-testing' the extent of moderation and terms of service limits in order to see whether they could use the platform to host their discussions, with high-profile early adopters including Nick Fuentes and Laura Loomer.[11] However, very little is understood about the nature and extent of TVE incursion beyond select journalistic accounts.[12]

Surface 2: Groups-based Functionality (e.g., Facebook Groups & Clubhouse Houses)

TVE actors have also become more creative with the groups that they join with the intent to become a trusted and engaging member of that online community. They join mainstream social political open or closed groups, even something as simple as "I Love Dogs," and post content or comments to spur any kind of discourse and interaction. In what platforms described as a 'public-private' strategy shift, extremists in these spaces are attempting to phish like-minded individuals based on responses and then move over to Messenger or another private venue to coordinate public-facing actions and/or vet potential recruits.

Surface 3: Direct Messaging Functionality (e.g., Facebook Messenger, Instagram Direct Messaging & Clubhouse Backchannel)

To engage in initial outreach to people who might be interested in extremist content but haven't fully made the leap, TVE actors have also used direct messaging functionality to post less extreme content. They are doing this to generate more one-to-one introductions with potential recruits but also for the

9 For more on this example, see Hugo Micheron, "Praising Jihadist Attacks on TikTok and the Challenge of Protecting Youths from Online Extremism," GNET Insight, December 9, 2020, https://gnet-research.org/2020/12/09/praising-jihadist-attacks-on-tiktok-and-the-challenge-of-protecting-youths-from-online-extremism/.

10 For more on this example, see: Gabriel Weimman and Natalie Nasri, "Hate on TikTok," GNET, July 7, 2020, https://gnet-research.org/2020/07/07/hate-on-tiktok/.

11 Zachary Petrizzo, "Clubhouse, popular new conversation app, starts booting far right extremists," Salon, May 12, 2021, https://www.salon.com/2021/05/12/clubhouse-popular-new-conversation-app-starts-booting-far right-extremists/.

12 For fairly scattered efforts on talking about TVE with relation to Clubhouse, see Dominik Hammer, Paula Matlach, Lea Gerster, and Till Baaken, "Escape Routes: How far right actors circumvent the Network Enforcement Act," Institute for Strategic Dialogue, October 22, 2022, https://www.isdglobal.org/wp-content/uploads/2022/11/escape-routes-how-far right-actors-circumvent-the-network-enforcement-act.pdf; Dawan M. Rohmatullah, "Digital Santri: The Traditionalist Response to the Religious Populism Wave in Indonesian Islam," Asian Studies, The Twelfth International Convention of Asia Scholars (ICAS 12), 1 (June 2022): 601 - 608, https://www.aup-online.com/content/papers/10.5117/9789048557820/ICAS.2022.069.

initial vetting of the potential candidate for both interest/susceptibility and to determine whether they are safe for the extremist organization to engage in further outreach.[13] There are also attempts to share personal usernames or contact information with the purpose of moving people "off-platform" in order to build community and engage in more extremist activity elsewhere.

Surface 4: Search Function (e.g., Clubhouse Explore & TikTok Discover)

Search is used for like-minded actor networks to connect with each other, which means it can be abused by TVE actors by using coded terms to evade detection or searching for risky but not necessarily violating groups or key terms. At a more basic level, extremists also use search to find users who are evading account bans and looking for previous accounts in order to reconnect with them. There are also attempts to share propaganda and coded hashtags via search results in order to evade content moderation, monitoring, detection, and disruption.[14]

Conclusion

NSM surfaces provide both an entry point to TVE exploitation but also P/CVE interventions. What creates a challenge in all of these surfaces when it comes to P/CVE efforts is the ephemerality, virality, and privacy of content, the difficulties of engaging in closed surfaces and spaces, the challenges of different media (e.g., content moderation on video and audio versus text) and the inability to take a one-size-fits-all approach. This is not unusual to P/CVE efforts on social media platforms historically but may require some retooling in order to better populate these surfaces with relevant, timely, and tailored interventions. In the next section, we will discuss how these challenges and barriers can be overcome through a set of discrete P/CVE interventions going forward.

Section II: Effective P/CVE Interventions on NSM Surfaces

Introduction

Over the past decade and a half, preventing and countering violent extremism initiatives have become a notable part of efforts to combat terrorism. Placed at the softer end of counter terror (CT) tactics, the use of resilience initiatives, counterspeech, deterrence or inoculation messaging, and search redirect in order to disrupt organizations committed to violent extremist causes has come to occupy the 'upstream,' 'midstream,' and 'downstream' spaces of preventative measures at the conceptual and behavioral level available to governments, Non-Governmental Organizations (NGOs) and civil society

13 For more on vetting, see the example of Fascist Forge in the following ISD report: Jacob Davey, Mackenzie Hart, and Cécile Guerin, "An Online Environmental Scan of Right-wing Extremism in Canada," Institute for Strategic Dialogue, 2020, https://www.isdglobal.org/wp-content/uploads/2020/06/An-Online-Environmental-Scan-of-Right-wing-Extremism-in-Canada-ISD.pdf.

14 For more on TVE usage of hashtags across different extremist ideologies, see Seth G. Jones, " The Rise of Far-Right Extremism in the United States," Center for Strategic and International Studies, November 2018, https://www.csis.org/analysis/rise-far right-extremism-united-states; Tina Nguyen and Mark Scott, "'Hashtags come to life': How online extremists fueled Wednesday's Capitol Hill insurrection," Politico, January 8, 2021, https://www.politico.eu/article/hashtags-come-to-life-how-online-extremists-fueled-wednesdays-capitol-hill-insurrection/; Lena Clever, Tim Schatto-Eckrodt, Nico Christoph Clever, and Lena Frischlich, "Behind Blue Skies: A Multimodal Automated Content Analysis of Islamic Extremist Propaganda on Instagram," Social Media + Society 9, no. 1 (2023), https://doi.org/10.1177/20563051221150404.

actors (see Figure 1 below). These have become especially important as terrorist organizations have become more adept at using social media to radicalize, recruit, and disseminate their messages online, thereby circumventing traditional media and face-to-face encounters. Our current moment therefore calls for a "war of words and ideas" as much as CT actions on the ground to combat the threat of extremist violence.

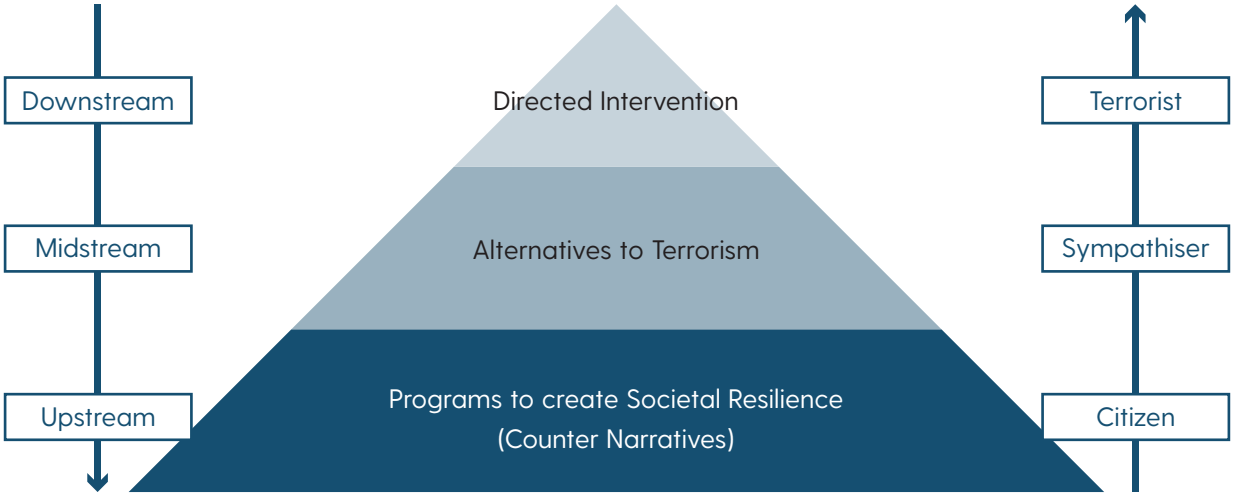


Figure 1: P/CVE within a Broader CT Strategy[15]

In this section of the playbook, we will identify the continuing utility of resilience initiatives, counterspeech, deterrence & inoculation messaging, and search redirect when it comes to these new and emerging surfaces. In what follows, we distinguish between holistic and targeted interventions – when referring to broader versus more targeted programming – as well as indicating the appropriateness of such interventions for upstream, midstream and downstream audiences (based on their level of radicalization and alignment with TVE actors).

(For more information on which interventions are best placed on what surfaces, please see Appendix).

Resiliency Initiatives (Holistic Upstream)

The first intervention approach envisaged is a ground-up partnership approach to interventions that equip local communities with the resources they need from more indirect offline actors (such as community leaders, local NGOs or key workers) to help off-ramp at-risk users and thus create resiliency among primary audiences.[16] The understanding is that the most significant impact is made at

15 Adapted from Anne Aly, Anne-Marie Balbi & Carmen Jacques, "Rethinking countering violent extremism: Implementing the role of civil society," Journal of Policing, Intelligence and Counter Terrorism 10, no. 1 (2015): 3-13, DOI: 10.1080/18335330.2015.1028772.

16 We define resiliency as online initiatives that are upstream (i.e., holistic) and have a substantive offline element to delivery (e.g., face-to-face educational initiatives). We also define resilience here as a form of social resistance (see Michele Grossman, "Resilience to Violent Extremism and Terrorism: A Multisystemic Analysis," Deakin University, January 1, 2021, https://hdl.handle.net/10536/DRO/DU:30156356) which – in a P/CVE arena – can mean both "withstand[ing] violent extremist ideologies" and also "challeng[ing] those who espouse them" ("Building Resilience Against Terrorism: Canada's Counter-terrorism Strategy," Public Safety Canada (2013): 11).

the local level, and platform staff are sometimes not best equipped to intervene in those contexts. The goal of this intervention approach would be to work with local partners and leverage platform profiles among communities and resources to enable local partners to do off-ramp work more efficiently in those contexts.

Several platforms already engage in such off-platform or organic content creation initiatives that might provide key insights for addressing the barriers within NSM regarding the ephemerality, virality, and privacy of content:

- **Meta:** Meta prioritizes where to target its resiliency interventions through available partnerships (for existing initiatives, see: https://counterspeech.fb.com/en/), existing metrics of prevention efforts, and known offline moments that might heighten the prevalence of hateful, extremist and terrorist use of the platform (e.g., high-risk elections and terror attacks). They admit it is a challenge to figure out what are the right areas to target, especially given the ephemerality of content on NSM spaces.
- **TikTok:** TikTok also takes a targeted approach to resiliency initiatives. It chooses regions based on internal metrics where the worst potential search terms are used and launches partnerships on the ground with NGOs to ensure the successful launch of campaigns. TikTok's development of partnerships has been strong in the Asian and Pacific markets, as there exists a high volume of easy-to-access and available partners with on the ground resources.
- **Clubhouse:** Clubhouse's resiliency partnerships to date have been driven chiefly by where it sees cognate forms of expertise emerging. For example, Clubhouse has partnered with organizations and individuals who study or are familiar with issues in Iran, Turkey, India, and Thailand. They also have partners in Europe and in the US related to combating hate speech and antisemitism.

Case Study: ***Facebook's Resiliency Initiative***

The Resiliency Initiative portal empowers local communities in the Asia-Pacific with digital tools to combat hate, violence, and conflict within and beyond their networks.[17] Launched in April 2021, this resource portal provides free access to tools in order to equip community networks in navigating the online space and use social media responsibly and effectively.[18] It also includes case studies in Bangladesh, Nepal, Sri Lanka, Malaysia and the Philippines of offline counter-prejudicial and counter-disinformation projects that have been used to tackle online harms through real-world interventions.[19]

17 See The Resiliency Initiative's Facebook Page: https://www.facebook.com/TheResiliencyInitiative/.

18 See The Resiliency Initiative's "About Us" Page: https://resiliencyinitiative.org/about/.

19 See The Resiliency Initiative's "Community" Page: https://resiliencyinitiative.org/community/.

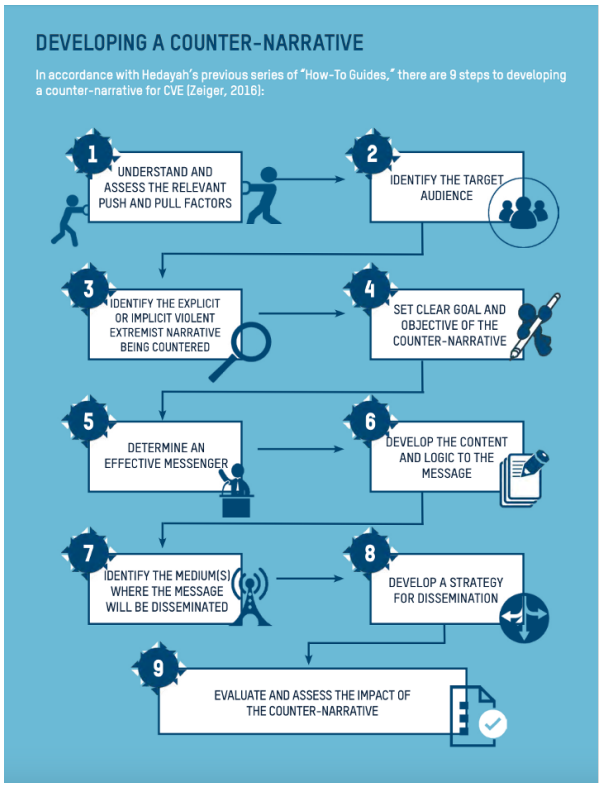


Figure 2: Developing a Counterspeech Campaign[21]

Counterspeech (Targeted & Holistic Upstream)

The second key intervention type intervention on NSM surfaces is counterspeech, defined as messages that can "[demystify], deconstruct or delegitimize extremist narratives."[20] Existing counterspeech interventions could be reused and remastered for NSM surfaces (see Figure 2 below for how such a campaign might be developed), but they would have to be heavily tailored, re-engineered and re-imagined to compete with the content focus of recommender algorithms.

In particular, counterspeech is highly contextual on NSM, and content moderation is complex and nuanced (and thus errors in moderation can occur). Such challenges can be met in several ways:

- First, platforms may wish to make it easier for creators to appeal moderation decisions on the platform itself and to invest in training for their Trust and Safety teams so they can better recognize counterspeech content apart from 'garden variety' terrorist and extremist content.
- Second, it is important for platforms and non-governmental actors to create an actual organic video and audio counterspeech campaign. Sometimes counterspeech campaigns can come across as advertisements or are highly edited, but such highly polished forms of counterspeech content do not necessarily perform better. Organic videos are the best and most effective, and a spontaneously created video is often more viral than a highly planned one. It might also be the case that alternative narratives (i.e., positive holistic messaging) work better than counterspeech (i.e., negative targeted messaging aimed at TVE narratives and tropes directly). This increases the need for NGOs and counterspeech activists to partner with platforms and their counterspeech

20 Henry Tuck and Tanya Silverman, "The Counter-Narrative Handbook," Institute for Strategic Dialogue, 2016, p. 65, https://www.isdglobal.org/wp-content/uploads/2016/06/Counter-narrative-Handbook_1.pdf.

21 For more see William Allchorn, "Building a Successful Radical Right Counter-Narrative Campaign: A How-To Guide," (Abu Dhabi, UAE: Hedayah, 2020).

services in order to streamline content.

- Finally, finding the right target audience is important but challenging. Because NSM surfaces are based on virality, ephemerality, and privacy (in addition to adversarial shifts), it can be harder to find very radicalized or very vulnerable individuals as they may not be very apparent to researchers and intervention providers (i.e., have private profiles or a network limited to a very niche community). It may therefore be challenging to use NSM surfaces to reach core audiences due to the nature of the product. Instead, videos supporting broader social change or promoting social good may be better suited for these surfaces. It might also be better to create a content community to influence those who might have access to vulnerable target audiences rather than trying to target those vulnerable audiences themselves.

Case Study: ***TikTok's Creator Support Campaign***

TikTok allows counterspeech and messages that undermine terrorist activity, convey sarcasm, provide educational or documentary content, or have significant scientific or artistic value without promoting terrorism. TikTok support has previously used hashtag and search tools to promote educational resources tagged with specific keywords (for example, to disseminate public service announcements related to sexual assault and to promote harm reduction resources and partnerships with NGOs and civil society.)

A key example of preventing violent extremism on TikTok is through their creator support campaign that ran from July to September 2022. This was a six-week "boot camp" to train, mentor, and support the creation of positive narratives on TikTok, and included creators aged 18-25 from Thailand, Indonesia, Malaysia and the Philippines, and was supported by the United Nations Development Programme and European Union (for examples, see the hashtag #creatorsupport on TikTok). A lot of the videos were created in local languages topics covered included: religious and ethnic tensions in Myanmar, religious tolerance in Indonesia, religious extremism in southern Thailand, and hate and polarisation in the Philippines.

Deterrence & Inoculation Messaging (Targeted Midstream & Downstream)

A third key intervention type intervention that can be deployed on NSM surfaces is deterrence & inoculation messaging. Deterrence messaging leverages the deterrence theory of punishment[22] and uses audience targeting (based on identifying audiences that are believed to be at risk due to their on-platform behavior but as yet do not meet the threshold for removal). For example, when a user logs on to Facebook or Instagram, they see a pop-up message from the platform that explains why Meta

22 The deterrence theory of punishment states that instead of being motivated by some deeper moral sense, people are deterred from committing crimes because they are afraid of getting caught. According to deterrence theory, people are most likely to be dissuaded from committing a crime if the punishment is swift, certain, and severe. For example, if there is a low likelihood that someone will get caught or if the punishment for getting caught is just a warning, deterrence theory says they will be more likely to steal it. For more exploration of this theory, see Hsin-Wen Lee, "Taking Deterrence Seriously: The Wide-Scope Deterrence Theory of Punishment," Criminal Justice Ethics 36, no. 1 (2017): 2-24, DOI: 10.1080/0731129X.2017.1298879.

is concerned with their platform behavior and provides links to on-platform civil society and NGO partner resources for further kinds of interventions depending on the specific concerns.

Such interventions could be added to those used on platforms already by acting as friction points before individuals view content or by providing attitudinal inoculation messaging for those further upstream before engaging with radicalizing networks or content. Such friction points already follow automated disinformation and misinformation efforts on platforms such as Twitter,[23] but instead of forcing users to read an article, it could be used to flag problematic content. In the latter sense, a warning could be served to the user suggesting an impending attempt to change beliefs or attitudes and an encouragement to click through to an article where a diluted version of extremist propaganda and counter-arguments are presented.

Case Study: ***Facebook Harm Prevention Initiatives***

Meta currently operates several pop-up messaging "nudges" in its P/CVE work on-platform that could be harnessed for NSM surfaces. The primary one is a so-called deterrence intervention where users with 3+ DOI Strikes – based on a mixture of TVE and content violation signals – are served a customized message (including their name) that informs them of platform standards and how their behavior is putting them at risk of a complete ban. They are also sent a Help Center article that helps individuals understand the rules on violating content better, the signs and dangers of radicalization, and how they can engage in more prosocial behavior on-platform. The secondary "nudge" is a so-called inoculation intervention that is used to reach users on the periphery of dangerous networks who are also indicating radicalization and reduce their propensity to do harm. This is targeted at midstream users vulnerable to radicalization who also have significant DOI behavioral signals with samples and filters built in to improve precision regarding who is targeted. In this case, a pop-up warning is served to the identified user that – in accordance with attitudinal inoculation theory[24] – suggests an impending attempt to change their beliefs or attitudes and encouragement to click through to a Help Center article where a diluted version of extremist propaganda and counterarguments are presented as well as support from NGOs that specialize in counter-radicalization counseling and DDR (i.e., disarmament, demobilization and reintegration) programming.

Search Redirect (Targeted Downstream)

A final key P/CVE intervention that was envisaged by the working group and can be harnessed for NSM surfaces is search redirect (often referred to as 'the Redirect Method'). Pioneered by Moonshot CVE, Jigsaw, and some larger online tech companies, search redirect is an initiative for when any actor is using the search surface. If they search for a keyword that is flagged as being a pathway

23 Andrew Hutchinson, "Twitter Shares Insights Into the Effectiveness of its New Prompts to Get Users to Read Content Before Retweeting," Social Media Today, September 24, 2020, https://www.socialmediatoday.com/news/twitter-shares-insights-into-the-effectiveness-of-its-new-prompts-to-get-us/585860/.

24 For more on attitudinal inoculation with regard to CVE, see Kurt Braddock, Weaponized Words: The Strategic Role of Persuasion in Violent Radicalization and Counter-Radicalization (Cambridge, UK: Cambridge University Press, 2020): Chapter 4.

linked to a terrorist or violent extremist group, they are surfaced in a module that informs them that their query might be associated with that harmful group. The module is offered on Meta products and uses partnerships with local NGOs to build out information and resources tailored to local needs and contexts in order to provide better counter-radicalization outcomes (e.g., lack of recidivism). Moreover, this is one of the most regularly deployed interventions throughout the world, with Google and YouTube offering similar initiatives in the US, UK, Australia, Germany, Pakistan, and Indonesia (and new launches planned for this year). When it comes to NSM surfaces, this approach could be further harnessed by serving counterspeech content or off-ramping individuals and giving them an opportunity to speak to a trained counselor.

Case Study: ***Institute for Strategic Dialogue Counter Conversations Pilot***

A large follow-on project from ISD's original "one-to-one pilot,"[25] ISD's "Counter Conversations" also used Facebook in order for former far right and Islamist extremists to communicate counter-narratives to 800 individuals showing clear signs of radicalization. This is one step up from counterspeech and posits a 'hybrid' model by challenging people one-to-one in order to help persuade individuals to exit online extremist milieus. Its results found that the approach was largely effective in sustaining conversations (71% of Islamist extremist interactions and 64% of radical right extremist interactions) and providing a lasting positive impact on the trajectory of individuals selected for the pilot (with 10% of individuals expressing an interest in taking their conversation offline, believing that their beliefs had been changed or challenged, and reducing their propensity to engage in negative online posting behavior).[26]

Conclusion and Recommendations

This playbook highlights the different NSM surfaces that have become key vectors for TVE exploitation on social media in recent years as well as how P/CVE interventions might be trained to meet this challenge. Though challenges exist in the ephemerality, virality, and privacy of content, the difficulties of engaging in closed surfaces and spaces, the challenge of different media (e.g., content moderation on video and audio versus text) and the inability to take a one-size-fits-all approach across different geographies, a number of recommendations can be made for policy makers, government officials, practitioners, researchers, and tech companies when exploring TVE exploitation of NSM surfaces and deploying P/CVE techniques going forward:

25 Ross Frenett and Moli Dow, "One to One Online Interventions – A Pilot CVE Methodology," Institute for Strategic Dialogue, September 2015, https://www.isdglobal.org/isd-publications/one-to-one-online-interventions-a-pilot-cve-methodology/.

26 Jacob Davey, Jonathan Birdwell, and Rebecca Skellett, "Counter Conversations: A model for direct engagement with individuals showing signs of radicalisation online," Institute for Strategic Dialogue, 2018, p. 7, https://www.isdglobal.org/wp-content/uploads/2018/03/Counter-Conversations_FINAL.pdf.

For Policy makers and Governments

Government Information & Communication Technology (ICT) policy reforms

A first key (and perhaps obvious) recommendation for government and policy makers is a call to update the metrics and objects of ICT policies to allow for the inclusion of NSM surfaces and platforms within their scope of interdiction, disruption, investigation, and legislation. More specifically, it might involve taking down hair triggers for the further exploration of TVE exploitation of such NSM surfaces and platforms as well as the opening of new lines of investigation in this regard. This could lead to further collaboration and consultation with NGOs and academics studying such new fields of exploitation and better harnessing existing tools to meet the threat.

Online harms related to NSM

A second key recommendation for government and policy makers is to take seriously the hateful extremism and soft recruitment efforts circulated by Non-Violent Extremist (NVE) and TVE actors on NSM surfaces. Governments should task government agencies and responsible government units with exploring the nature and scope of the TVE threat in these spaces and the potential online harms that might result from interaction with such content. In particular, it should include pressure for tech platforms and policy makers to act when such content is found.

Training and Protections for Influencers

Another key recommendation for government and policy makers – as well as platforms and NGOs – is that if influencers are increasingly used to front counterspeech and resiliency initiatives, then they should be provided with the training and protections needed to make sure they know the topic that they are speaking on and how they can seek help if TVE communities decide to enact coordinated online targeted harassment against them. This will allow them the creative space to design the intervention in a voice and style suited to them.

Identification of local actors for resiliency initiatives

One final role that governments and policy makers can take in NSM P/CVE programming is identifying local public and private actors that might be best placed to enact broad-based resiliency initiatives. Governments have the tools to identify community leaders, local NGOs or key workers that might be best placed to run 'offline' educational, counterspeech and pre-bunking campaigns at the local community level. They also have the funding to enact such projects – tailored to emerging threats and needs – and therefore stop radicalization before it arrives in the online space.

For Platforms

Transparency from tech companies

A first (and perhaps perennial) recommendation for tech companies when it comes to NSM surfaces is transparency with researchers and practitioners about what their monitoring and disruption efforts show in terms of NSM and what they are doing about it. This would go a long way to brief government,

NGO, and practitioner partners on the types of threats they are seeing and therefore tailor their research and interventions at key inflection points. Such a recommendation is therefore a call for further co-creation in this space, using roundtables, events, and briefings to further sharpen each other's work for positive real-world good.

Tech company incentives related to affordance for NSM P/CVE efforts

As new campaigns emerge on NSM surfaces, tech companies need to respond with in-kind and sponsored support of such content. This can spring from existing counterspeech and blue-teaming efforts but might require smaller platforms to devote greater (and often scarce) Trust and Safety resources to countering and preventing the threat posed by TVE exploration on-platform. Meta and TikTok already operate Safety Ads and Creator Support programs, but programming focused on NSM surfaces is needed to see whether these innovations can be used to mitigate exploitation by bad actors.

Resources surrounding the Monitoring, Measuring & Evaluation (MME) of interventions to ensure impacts are effective and not negative

Due to the increasing movement of platforms toward content-based algorithmic distribution of (mainly video) content, wider distribution of how P/CVE practitioners and civil society organizations can measure, monitor, and evaluate blue-teaming efforts on platforms would considerably help in making sure research and intervention efforts are properly tailored and attuned to these emerging dynamics. GIFCT and several platforms already provide guidance on MME in this space (e.g., Campaign Toolkit and FB Counterspeech), but a greater array of on-platform metrics and resources on MME would certainly help researchers and NGOs alike evaluate their testing efforts.

Guardrails against Potential TVE Backlash to countermeasures

Due to the reflexive and reactive nature of TVE actors to various interventions and countermeasures enacted on platforms, it is recommended that platforms and intervention providers 'red-team' potential negative responses. For counterspeech, this might involve active monitoring against 'raids' or preventing counter campaigns as early as possible, as well as A/B testing content among focus groups that are as close to or representative of the audience in order to forestall so-called 'backfire' effects. For deterrence warnings, it might also include (where possible) using IP tracking, artificial intelligence, hashing, or other technological approaches to detect when such warnings are being exploited by TVE actors to recruit and propagandize within their support bases. Such guardrails will then help to mitigate potential harms or adversarial behavior caused by prescribed intervention approaches – adhering to a 'do no harm' principle as a universal standard in P/CVE.

For NGOs & Researchers

Further research into TVE use of these platforms

In order to better counter the threat of TVE recruitment and radicalization on NSM surfaces, increased funding for more research on the nature and scope of extremist exploitation is recommended. While

a few promising studies have emerged, more systematic investigation of public and private dynamics, the use of coded and ephemeral content, and phishing attempts by TVE actors in 'normie' territory need to be better delineated to meet the threat head-on. This goes for TikTok and Clubhouse in particular as well as the use of audio as a new and separate medium of TVE exploitation.

More research on what an effective intervention looks like for different target audiences with a focus on long-term impact

Due to the emerging nature of NSM P/CVE, funding for additional research into the nature and scope of what constitutes ethical and effective P/CVE interventions in this new space of ephemerality, virality, and privacy is recommended. While researchers and NGOs might be able to riff off of the lessons and previous studies of on-platform engagement with regards to these themes and what safeguards need to be met when it comes to interventions, more specific experimental efforts that show and test awareness, engagement and impact of a whole suite of retooled interventions need to be broached in order to see whether they are effective and are doing 'no harm' when it comes to the target audience.

Literacy on-platform style championed by academia and tech NGOs

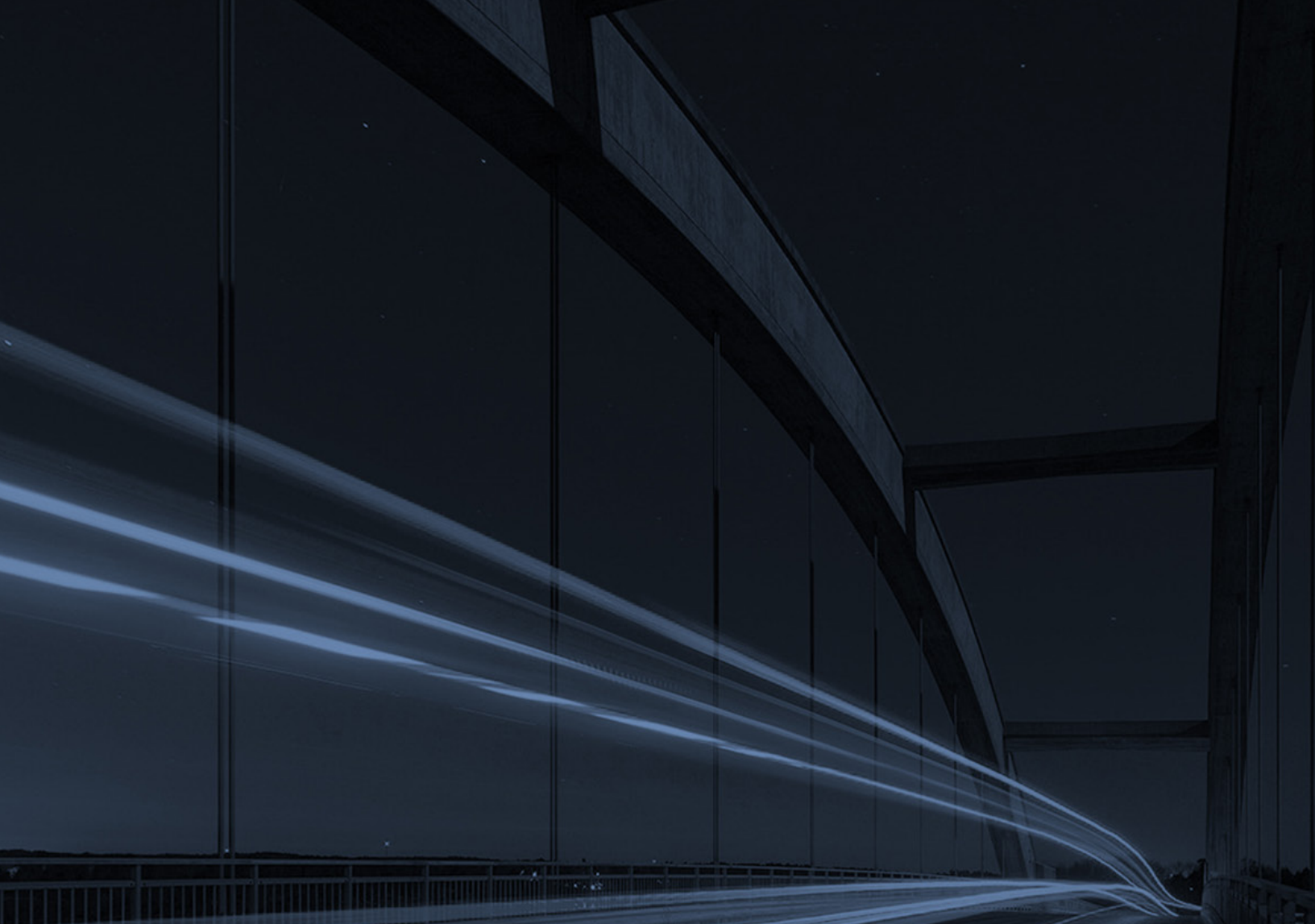
Due to the increasing movement of platforms toward algorithmic distribution of (mainly video) content, more widely distributed information on how P/CVE researchers and civil society organizations can exploit this shift for blue-teaming purposes by platforms would go a great way in making sure research and interventions efforts are properly tailored and attuned to these new and emerging dynamics. GIFCT and several platforms already provide guidance in this space passively (e.g., Campaign Toolkit and Facebook's Counterspeech resources), but greater awareness and engagement actively by such entities would help researchers and NGOs alike.

Partnerships between tech and NGO organizations to provide links to outreach companies for individuals targeted by intervention.

Though this is happening already with regard to resiliency initiatives, search redirect, counterspeech, deterrence and inoculation messaging, the potentially voluminous and disruptive nature of new upstream and midstream interventions means that more NGO partners will need to be onboarded to fulfill follow-up functions needed to make sure that individuals targeted by such interventions are properly debriefed. This taps into the 'do no harm' principle of CVE but also makes sure that further 'backfire' effects of such interventions are contained and managed in an ethical and responsible way.

Appendix: NSM Surfaces vs. P/CVE Interventions Matrix

Surfaces (Y) vs. Intervention (X)	Resiliency Initiatives (Holistic Upstream)	Counterspeech (Targeted & Holistic Upstream)	Deterrence & Inoculation Messaging (Targeted Midstream & Downstream)	Search Redirect (Targeted Downstream)
TikTok Search				X
TikTok's For You Feed	X	X	X	
TikTok Live Stream		X		
Clubhouse Homepage	X		X	
Clubhouse Search				X
Clubhouse Live Stream		X		
Clubhouse Rooms	X			
Clubhouse Events	X			
Clubhouse Clubs	X	X		
Clubhouse Backchannel		X		X
Clubhouse Explore				X
Facebook Search				X
Facebook Newsfeed	X	X	X	
Facebook Timeline	X		X	
Facebook Messenger	X	X	X	X
Facebook Stories		X		
Facebook Live		X		
Facebook Groups	X			
Instagram Feed	X		X	
Instagram Group Profile	X			
Instagram Reels		X		
Instagram Stories		X		
Instagram Videos		X		
Instagram Direct Messaging	X	X	X	X
Instagram Search				X
Instagram Live		X		



Recommended citation: Dr. William Allchorn, Richmond, the American International University in London, Nouveau Social Media Playbook (Washington, D.C.: Global Internet Forum to Counter Terrorism, 2023), Year 3 Working Groups.

GIFCT is a 501(c)(3) non-profit organization and tech-led initiative with over 20 member tech companies offering unique settings for diverse stakeholders to identify and solve the most complex global challenges at the intersection of terrorism and technology. GIFCT's mission is to prevent terrorists and violent extremists from exploiting digital platforms through our vision of a world in which the technology sector marshals its collective creativity and capacity to render terrorists and violent extremists ineffective online. In every aspect of our work, we aim to be transparent, inclusive, and respectful of the fundamental and universal human rights that terrorists and violent extremists seek to undermine.

www.gifct.org **outreach@gifct.org**