

Playbook on Positive Intervention Strategies Online

GIFCT Blue Team Working Group

2022-2023



GIFCT
Global Internet Forum
to Counter Terrorism

Introducing GIFCT Year 3 Working Group Outputs

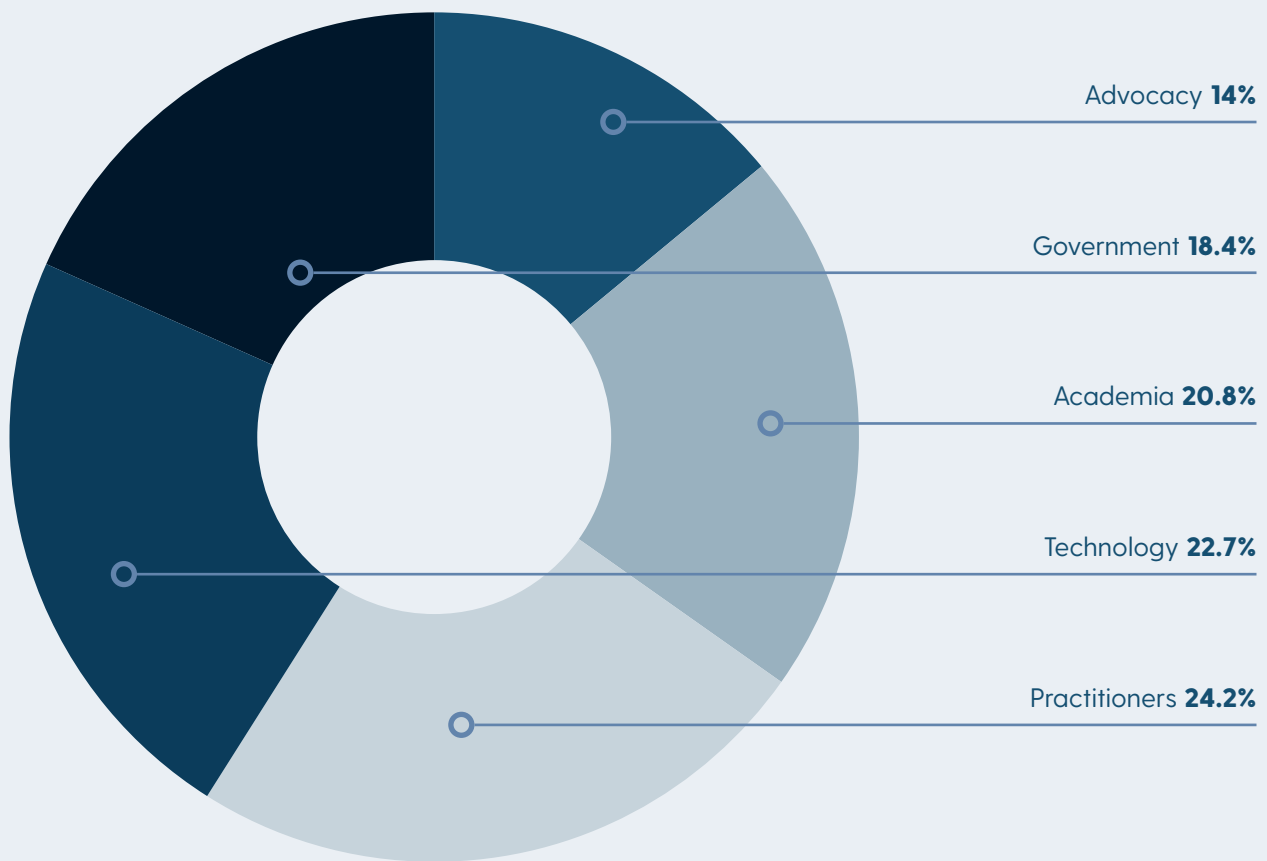
By Dr. Nagham El Karhili, Programming and Partnerships Lead, GIFCT

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. Initiated 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](#).

Table of Contents

Introducing the Playbook by Dr. Erin Saltman and Scott Johnson	1
Nouveau Social Media by Will Allchorn	2
Gaming & Gaming-Adjacent Platforms by Samantha Kutner	16
Positive Interventions on Marketplace Platforms by Ellie Rogers	23
Strategizing Online Interventions In Global Contexts by Prof. Fredrick Ogenga	37

Introducing the Playbook

Dr. Erin Saltman and Scott Johnson, GIFCT BTWG Leads

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. The BTWG developed this Playbook to highlight alternative platforms for potential positive interventions. The result is a tailored set of approaches and best practices to further PVE/CVE efforts spanning a wider diversity of platforms. It aims to help activists in their efforts to challenge hate and extremism online and foster broader CSO-Tech Company partnerships looking at intervention potentials on (1) Nouveau Social Media platforms, (2) gaming platforms, and (3) lifestyle and marketplace platforms. It also includes a chapter on (4) regional and cultural sensitivities for positive interventions. All Working Group outputs are made available on the [GIFCT Working Groups page](#).

The types of platforms discussed in each session were chosen by group participants comprised of cross-sector experts. The BTWG had a diversity of international practitioners and individuals experienced in funding, developing, and launching online PVE/CVE interventions. Each section of the Playbook was led by a Working Group participant who synthesized the group's dialogues from the sessions alongside research and interviews with tech companies. GIFCT invited platform representatives relevant to each platform theme to attend sessions and present to the group. Discussions included questions about the viability of positive interventions and considerations activists and practitioners should take alongside existing resources that could facilitate PVE/CVE approaches. While the group attempted to find replicable and scalable best practices and advice, there was also the recognition that interventions need local nuance and understanding to develop impactful content. The final section of the Playbook highlights the need for localized partnerships and expertise to sit alongside platform or internationally-led intervention approaches.

This Playbook builds off of GIFCT's previous Working Groups. In 2021, the Working Group focused on *Content-Sharing Algorithms, Processes, and Positive Interventions*. This group [developed a paper mapping online intervention strategies, theoretical frameworks, and tactics](#), including case studies of real-world PVE/CVE intervention examples. In 2022, the Working Group on *Positive Interventions and Strategic Communications* developed both a [practitioner's guide for measuring impact and audience engagement](#) and a [guide on good practices, tools, and safety measures](#) for researchers and practitioners focused on terrorist and violent extremist groups.

GIFCT hopes that this Playbook and previous Working Group outputs can help inspire civil society, government, and tech company teams involved in PVE/CVE work to innovate and explore a broader range of positive intervention approaches online.

Nouveau Social Media

Will Allchorn, Richmond American University London

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.¹ 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.² 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,³ the use of manifestos, online meme culture, and conspiracy theories can lead to powerful offline effects,⁴ with the seemingly sporadic and “solo” actor nature of the attacks masking a broader toxic online network of religiously, ethnically, and racially motivated hatred.⁵

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 are defining here as ‘Nouveau Social Media’ (NSM).⁶ This output 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

1 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.

2 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/>.

3 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 Roma, “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-extremism-into-focus/>.

4 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>.

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

6 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.

survey of existing Nouveau Social Media P/CVE efforts, this section of the Playbook helps define recent adversarial shifts and possible effective interventions on these types of surfaces. We will then conclude by proposing possible recommendations for scientifically-proven, sustainable, and scalable P/CVE efforts that could be applied elsewhere.

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.⁷ 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.

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.⁸ 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).⁹ 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

7 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).

8 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/>.

9 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/>.

adopters including Nick Fuentes and Laura Loomer.¹⁰ However, very little is understood about the nature and extent of TVE incursion beyond select journalistic accounts.¹¹

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 trusted and engaging members 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 have not 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 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.¹² 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 TVE actors can abuse it 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,

10 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/>.

11 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>.

12 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>.

monitoring, detection, and disruption.¹³

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. Next, we will discuss how these challenges and barriers can be overcome through a set of discrete P/CVE interventions going forward.

Effective P/CVE Interventions on NSM Surfaces

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 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. Therefore, our current moment calls for a "war of words and ideas" as much as CT actions to combat the threat of extremist violence.

.....

13 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>.



Figure 1: P/CVE within a Broader CT Strategy¹⁴

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.¹⁵ The understanding is that the most significant impact is made at the local level, and platform staff are sometimes not best equipped to intervene in those contexts. This intervention approach aims 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

¹⁴ 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.

¹⁵ 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).

(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 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 is 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 the U.S. 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.¹⁶



Launched in April 2021, this resource portal provides free access to tools to equip community networks to navigate the online space and use social media responsibly and effectively.¹⁷ 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.¹⁸

Counterspeech (Targeted & Holistic Upstream)

The second key intervention type on NSM surfaces is counterspeech, defined as messages that can "[demystify], deconstruct or delegitimize extremist narratives."¹⁹ 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.

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

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

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

19 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.

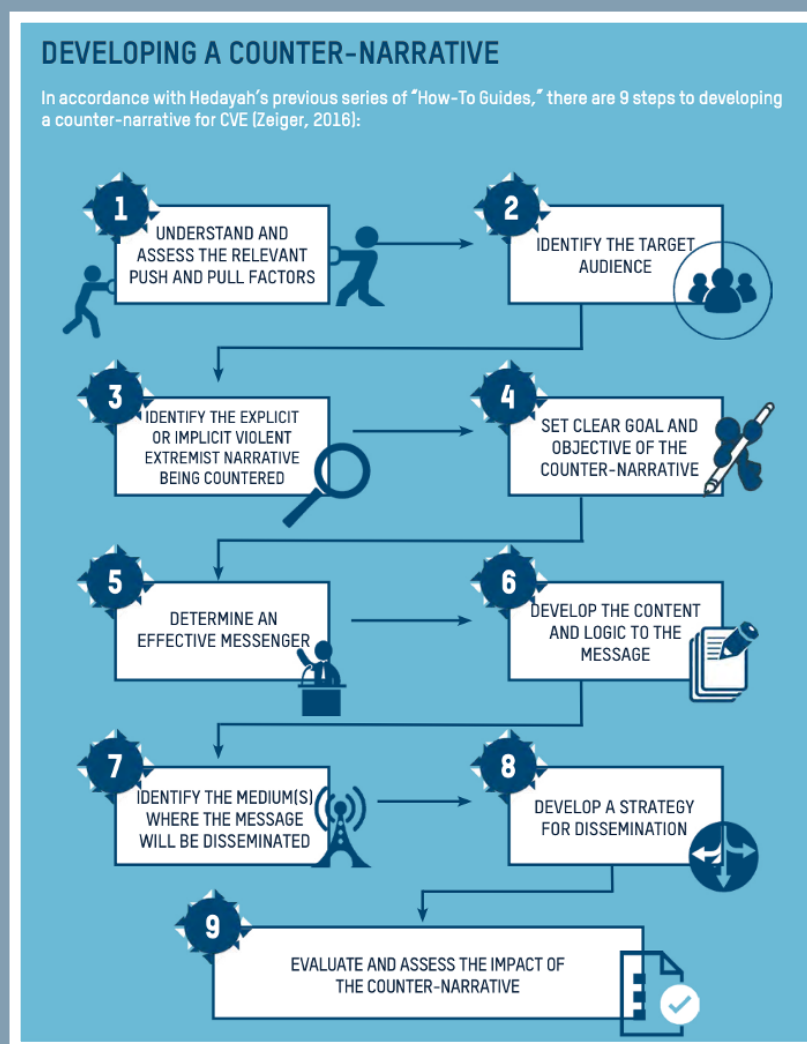


Figure 2: Developing a Counterspeech Campaign²⁰

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

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

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 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 example, see the hashtag #creatorsupport on TikTok). Many videos were created in local languages and covered religious and ethnic tensions in Myanmar, religious tolerance in Indonesia, religious extremism in southern Thailand, and hate and polarization 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 punishment²¹ and

.....

²¹ 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.

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 explaining why Meta is concerned with their platform behavior and providing links to on-platform civil society and NGO partner resources for further 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,²² 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 counterarguments 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 better understand the rules on violating content, the signs and dangers of radicalization, and how they can engage in prosocial behavior on-platform. The secondary “nudge” is a so-called inoculation intervention used to reach users on the periphery of dangerous networks who also indicate 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 theory²³ – 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

.....
 22 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/>.

23 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.

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 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 U.S., 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,"²⁴ 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).²⁵

Conclusion and Recommendations

This output 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:

24 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/>.

25 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 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 prebunking 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 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, and 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 actively 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 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 TVE actors are exploiting such warnings 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, coded and ephemeral content use, 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 active awareness and engagement 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 ensures that further 'backfire' effects of such interventions are contained and managed ethically and responsibly.

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		

Gaming & Gaming-Adjacent Platforms

Samantha Kutner, Glitterpill LLC

Introduction

In recent years, the popularity of online video games has skyrocketed. Revenue from the market for video games is projected to reach \$385 billion in 2023, with a year-on-year growth of 7.9% until 2027.²⁶ As online communities continue to grow and new payment integrations become available, ensuring the safety of all users in these spaces and anticipating potential threats has become increasingly urgent.

This growth is occurring in parallel with drastic cuts to trust and safety teams across the tech industry, which can undo much of the work done to protect users.²⁷ The exponential growth of gaming and gaming-adjacent platforms has outpaced companies' willingness to keep online spaces safe. Without increasing the capacity of trust and safety teams, users may feel uncomfortable or even unsafe using a platform.²⁸ This can lead to decreased engagement and ultimately harm the platform's reputation and success.

This report discusses ways to prevent harmful behavior in these spaces, address online safety issues, and offer positive interventions for gaming and gaming-adjacent platforms. Glitterpill's Chief Intelligence Officer heard from multiple Blue Team Working Group members, practitioners, experts, parents, and even teen users. Building on this, we provide best practices for each intervention stage, along with additional considerations for positive interventions.

TVEC Exploitation on Gaming and Gaming-Adjacent Platforms

Recent headlines from platform spaces reveal the hectic pace of expansion:

- Major platforms like Epic are starting to build out their platforms in the metaverse.²⁹
- On July 12th, Meta announced that Roblox will soon be added to its expanding VR gaming push.³⁰

26 "Video Games - Worldwide: Statista Market Forecast," Statista, accessed August 1, 2023, <https://www.statista.com/outlook/dmo/digital-media/video-games/worldwide>.

27 Alex Heath, "Facebook Splits up Unit at Center of Contested Election Decisions," The Information, December 3, 2020, <https://www.theinformation.com/articles/facebook-splits-up-unit-at-center-of-contested-election-decision>.

28 Isabella Zavarise, "Migration to Other Social Media Platforms Shows No Signs of Slowing Following Elon Musk's Chaotic Takeover at Twitter, Report Says," Business Insider, December 6, 2022, <https://www.businessinsider.com/twitter-migration-shows-no-signs-of-slowing-following-musks-takeover>.

29 Jay Peters, "Tim Sweeney Wants Epic to Help Build a Metaverse That's Actually Positive," The Verge, December 15, 2022, <https://www.theverge.com/2022/12/15/23511494/tim-sweeney-epic-games-metaverses-positive-dystopian>.

30 <https://www.meta.com/en-gb/blog/quest/roblox-open-beta-app-lab/>; Andrew Hutchinson, "Meta Establishes New Partnership to Bring Roblox to Quest VR," Social Media Today, July 12, 2023, <https://www.socialmediatoday.com/news/meta-establishes-new-partnership-bring-roblox-to-quest-vr/686826/>.

- As of July 26th, 2023, Xbox will accept Venmo payments in games.³¹

However, with expansion comes risk. Extremist groups have been known to exploit gaming and gaming-adjacent platforms for various purposes, including recruitment, crowdfunding, mass harassment, brigading, and stochastic terrorism campaigns.³² They and other right-wing extremist groups and individuals across the globe have made use of live-streaming platforms such as Twitch, YouTube Live/YouTube Gaming, Facebook Watch/Facebook Gaming, Instagram Live, TikTok Live, Younow, DLive, Trovo, and Steam, as well as discussion platforms and forums like Discord, Reddit (r/gaming), IGN Boards, Minecraft Forum, GameFAQ, Steam, 4Chan, 8Kun, and Kiwifarms.

In some cases, individuals have manipulated the media and spread disinformation, such as frequently falsely identifying comedian Sam Hyde as the gunman in mass shootings.³³ Recruiters have used gaming references to frame their calls to action, providing weapons and access to bomb-making materials through gaming platforms. In one case, analysts found Discord users discussing how to evade new community guidelines with more innocuous references. In response to the rollout of new policies from the platform, these users also strategized how to pull individuals into different channels and platforms.

Extremists have discussed new ways to redpill potential recruits on a number of platforms and even recreated on Minecraft and Roblox³⁴ the 2019 neo-Nazi terrorist attack on Christchurch mosques. Recently, Patriot Front and Vanguard Britannia have used Roblox to hold virtual fight clubs and rallies.

Preventing TVEC Exploitation on Gaming & Gaming-Adjacent Platforms

It is crucial to address the exploitation of gaming and gaming-adjacent platforms by extremist groups. However, it is equally important to consider the perspectives of parents, teens, practitioners, and experts on radicalization in gaming-adjacent spaces as they work to prevent and counter violent extremism and develop human-centered solutions.

.....
31 Tom Warren, "You Can Now Venmo New Xbox Games," The Verge, July 26, 2023, <https://www.theverge.com/2023/7/26/23808367/venmo-xbox-microsoft-store-support>.

32 "Extremists' Use of Gaming (Adjacent) Platforms – Insights Regarding Primary and Secondary Prevention Measures, August 2021," n.d., Migration and Home Affairs, https://home-affairs.ec.europa.eu/networks/radicalisation-awareness-network-ran/publications/extremists-use-gaming-adjacent-platforms-insights-regarding-primary-and-secondary-prevention_en; "Gamers Who Hate: An Introduction to ISD's Gaming and Extremism Series – ISD," Institute of Strategic Dialogue, November 29, 2021, <https://www.isdglobal.org/isd-publications/gamers-who-hate-an-introduction-to-isds-gaming-and-extremism-series/>; "RAN C&N Digital Grooming Tactics on Video Gaming & Video Gaming Adjacent Platforms: Threats and Opportunities, Online Meeting 15-16 March 2021," n.d., Migration and Home Affairs, https://home-affairs.ec.europa.eu/networks/radicalisation-awareness-network-ran/publications/ran-cn-digital-grooming-tactics-video-gaming-video-gaming-adjacent-platforms-threats-and_en; "Press Release: Our Contribution to the January 6th Select Committee – The Khalifa Ihler Institute," The Khalifa Ihler Institute, July 12, 2022, <https://www.khalifahiler.org/newsb/2022/7/12/press-release-our-contribution-to-the-january-6th-select-committee>.

33 See <https://twitter.com/ashkenaz89/status/1342614521874018304?s=20>.

34 Mark Townsend, "How Far Right Uses Video Games and Tech to Lure and Radicalise Teenage Recruits," The Guardian, February 14, 2021, <https://www.theguardian.com/world/2021/feb/14/how-far-right-uses-video-games-tech-lure-radicalise-teenage-recruits-white-supremacists>.

Key Findings With Parents and Teens

In interviews, one Glitterpill analyst and parent emphasizes that when it comes to radicalization, “We’re not even talking about white nationalism. We’re just talking about the challenges that teenagers face on these platforms and how to keep themselves safe.”³⁵ She went on to observe that “if you want to talk about radicalization of white nationalists, you also at the same time have to talk about child predators that are on platforms.” She believes preventing child exploitation more broadly will have trickle-down effects because radicalization is also a form of grooming.³⁶

One Glitterpill analyst’s daughter is a 14-year-old girl who has had an active Discord account for two years. She described her experiences of being nonconsensually added to porn chats and suggested the following to enhance teen safety:

- Increased bot moderation and explicit gathering of consent before being added to groups or channels;
- Create a minimally invasive way for parents to monitor their child’s activity without needing to log into their account and review all content (parents could see the number of screen hours and how long their child has interacted with certain individuals, the minor could retain some sense of agency, and parent-driven discussions could be facilitated);
- Reconceptualize the notion of “kids being kids” and better understand how bullying can (in some cases) build rapport within the context of an ingroup to allow children more autonomy in gaming and gaming-adjacent spaces with their friends.³⁷

Key Findings With Experts

Galen Lamphere-Englund and Jessica White note how far-right actors “[exploit] historical gameplay to create alternative realities and downloadable game content based on white supremacist ideologies.”³⁸ This can promote harmful and extremist ideologies within gaming communities, which (if not addressed properly) can ultimately result in PR crises for gaming platforms. By combating historical revisionism through education and fact-checking, gaming platforms can create a safer and more inclusive environment for all users.

Lamphere-Englund, the Preventing and Countering Violent Extremism (PCVE) Advisor of Extremism And Gaming Research Network (EGRN), argues that “pushing developers to maintain historical accuracy”

.....
35 Personal Correspondence, July 17th, 2023.

36 “RAN C&N Digital Grooming Tactics.”

37 Daniel Koehler, Verena Fiebig, and Irina Jugl, “From Gaming to Hating: Extreme: Right Ideological Indoctrination and Mobilization for Violence of Children on Online Gaming Platforms,” *Political Psychology* 44, no. 2 (2022): 419–34, <https://doi.org/10.1111/pops.12855>.

38 Galen Lamphere-Englund, and Jessica White, “The Online Gaming Ecosystem: Assessing Socialisation, Digital Harms, and Extremism Mitigation Efforts,” *Global Network on Extremism and Technology (GNET)*, May 2023, <https://doi.org/10.18742/pub01-133>.

is not a constructive way to address resistance from game developers who may not see the value in maintaining historical accuracy. “Creative liberty and freedom of expression are key to making good games. Loose interpretations of history are rife, from Assassin’s Creed to Wolfenstein, and make for fun games.”³⁹ However, extremist repurposing of alt histories are another matter. To counter those, game producers need to become aware of how specific narratives in games, such as those around Vikings, historical religious groups, or cults can be exploited by extremist actors. Such games can be thoughtful in their messaging and even humorous in their rebuttals.

Game producers should also be aware which narratives might be particularly sensitive or play into extremist or terrorist recruitment tactics. Lamphere-Englund noted that “This is where collaboration with industry and researchers matters most: GIFCT, EGRN, and TAT could provide free or affordable resources to help train devs and studios on these risks along with sensible ways to mitigate them.”⁴⁰ Nevertheless, he reminded researchers that games are primarily “an escape for some, or a social world”:

Lectures – or games as edutainment – are good for classroom or online educational settings, but games need to be fun first, especially if being used as interventions.

He believes that “without a clear target audience and dissemination mechanisms (such as partnerships with educators or the like),” the likelihood of games functioning as successful intervention mechanisms is low.

Educational components that include friction points and historically accurate resources that users can click through have the potential to facilitate engagement. Researchers can suggest these components to appeal to their audience’s intelligence and allow them to learn about the inconsistencies in what they are being taught and how it does not accurately reflect history or reality.

A case study in October 2020 showed how Facebook redirected users searching for QAnon content to credible sources in order to “inform them of the real-world realities of QAnon and its ties to violence and real-world harm.”⁴¹ This initiative could be applied to combat historical revisionism in gaming with both upstream and downstream components. From an upstream perspective, when registering for a game, users could be required to engage with short, accessible, bite-sized educational content on historical revisionism. Downstream, they could be provided educational resources to combat historical revisionism in gaming, along with entertaining, fact-checked content from history buffs who create engaging, accurate content without embedded extremist narratives.

.....

39 Jennifer Locke, “Assassin’s Creed History: The Full Story (so Far),” Windows Central, June 2, 2022, <https://www.windowscentral.com/assassins-creed-story-so-far>; see also <https://store.epicgames.com/en-US/p/wolfenstein-the-new-order>; Galen Lamphere-Englund, “Home - Galen Lamphere-Englund - Englund Consulting,” January 4, 2023, <http://englundconsulting.org/>.

40 Personal correspondence, August, 2023.

41 “An Update to How We Address Movements and Organizations Tied to Violence,” Meta, October 17, 2022, <https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/>.

Key Findings With Practitioners

Blue Team Working Group members suggested Microsoft intervene in the game development space to educate developers. Additional suggestions involve having a youth reference group test content on a localized level, while another focus group could seek to understand the unique needs of neurodivergent audiences who may be more susceptible to gaming addiction.⁴²

Interventions to enhance community health range from punitive measures such as removing organizations to prioritizing user experience and introducing friction in the developer stage prior to product launches. Red teaming exercises with experts embedded in gaming and related fields could help enhance user safety and reach intended audiences while minimizing backfire. Proper expertise and smaller companies adopting a cultural shift towards prevention could signal to larger companies how to compete in the user safety realm.

A group of experienced practitioners specializing in facilitating disengagement completed a survey created to critically examine existing methods and techniques and assess their transparency and creativity. Practitioners provided detailed answers to a series of questions (see Appendix), elaborating on their experiences and providing insight into their unique approaches to disengagement facilitation.⁴³ Their responses were used to inform the following best practices for positive interventions in gaming-adjacent spaces.

Best Practices Throughout the Intervention Process

Before an Intervention

Deradicalization tends to be rooted in cognitive schemas, which can never be 100 percent accurately inferred. For this reason, targeting behavioral-based outcomes is preferable. In addition to enhancing user design and gaining a deeper understanding of target audiences prior to an intervention, it is vital to clarify research questions and tie them to overt measurable outcomes.⁴⁴

For example, instead of investigating ways to reduce the prevalence of alt histories in gaming spaces, researchers could ask, "Does probing historical inconsistencies with educational content from trusted influencers reduce the number of online raids on marginalized communities on [x platform]?" Similarly, instead of speaking about the need for privacy, researchers could ask, "Does enhancing privacy settings

42 Temma Ehrenfeld, "What Science Says about Video Games and ADHD," Healthline, September 16, 2018, <https://www.healthline.com/health-news/the-link-between-adhd-and-video-games#:~:text=The%20bottom%20line-,While%20video%20games%20do%20not%20cause%20ADHD%2C%20they%20can%20exacerbate,can%20lead%20to%20positive%20results>.

43 Identifying information about companies and programs have been removed.

44 Samantha Kutner, "The Call for Component Analyses of the Saudi Arabian Risk Reduction Initiative: An Examination of Religious Re-Education's Role in the Deradicalization and Disengagement Process," *Journal for Deradicalization* 7 (2016): 107–123, <http://scholarworks.unr.edu:8080/bitstream/handle/11714/990/Kutner%20-%202016%20-%20The%20Call%20for%20Component%20Analyses%20of%20the%20Saudi%20Arabi.pdf?sequence=1&is-Allowed=y>.

with teen accounts reduce the frequency of being nonconsensually added to porn groups and other harmful online spaces?”

During an Intervention

Practitioners emphasized the success of Exit USA programs that introduced friction points and provided links to organizations during sessions, reducing the response effort involved for at-risk individuals. Practitioners also emphasized the importance of enhancing digital literacy. This approach offers new opportunities for adapting to gaming platform interventions. Another suggestion was to encourage a general gaming audience to report problematic behavior as they encounter it on their preferred platforms (with a caution that this may result in false positives and the weaponization of the reporting function as described in prior ethnographic research).⁴⁵

After an Intervention

We have highlighted the importance of continued monitoring after major network disruptions. This aligns with Jigsaw’s recent assessment of the resurgence of the Boogaloo movement on mainstream platforms.⁴⁶ One of the most valuable survey questions involved honest reflection about program successes and acknowledging learning opportunities. It highlighted that the PCVE industry’s pressure to secure limited funding may lead organizations to celebrate successes when there are not any. When positive interventions have unintended consequences, it is crucial to retool them and address existing issues before continuing to disseminate and promote them. Transparency and accountability can help mitigate reputational damage and unintended harms.

Conclusion

Positive interventions on gaming and gaming-adjacent platforms are an important step in keeping online spaces safe. Increasing cooperation between PCVE practitioners like GIFCT, TAT, Glitterpill LLC, Moonshot CVE, ADL, Exit USA, Jigsaw, and PERIL can ensure best practices are adhered to at each stage of platform development. It is also crucial for organizations to offer realistic promises and provide truthful and transparent assessments of interventions within given constraints. All those involved in PCVE benefit from understanding that public trust is a form of currency as well.

.....
45 Samantha Kutner, “I Misgendered Alex McNabb’s Fake Twitter Account,” YouTube, 2019, <https://www.youtube.com/watch?v=dJRzRhazhVQ>.

46 “Facebook Banned ‘Boogaloo’-Related Groups — But New Research Suggests a ‘Boomerang’ Effect,” NBC News, July 14, 2023, <https://www.nbc-news.com/tech/internet/facebook-banned-boogaloo-groups-new-research-rcna93424>.

Appendix

Survey Questions

Q1: Describe an instance of an intervention that was intended to have a positive impact but was ultimately unsuccessful. What factors do you believe contributed to its failure? What might you or others have done differently?

Q2: What is something that could be added to countering violent extremism efforts on gaming platforms?

Q3: Describe an intervention that had positive outcomes. What factors do you think contributed to its success?

Q4: Describe something you are hesitant to try because you are not sure if it will work or not.

Condensed Survey Results

	Q1	Q2	Q3	Q4
Participant 1	Lack of a clear targeting strategy.	Data. It is not possible to mount a proportional response to a threat without understanding the scale.	Intervention: Individualized approaches to counseling, mentoring, and job placements for at-risk individuals, increasing sense of agency.	Measurement, monitoring, and evaluation of these interventions is crucial for them to be successful and minimize risk of harm. I am unsure how this could be carried out/ how counterspeech could be shared on gaming platforms, given they do not have typical timelines/ news feeds like mainstream traditional social media platforms.
Participant 2	Uncredible messenger, lack of campaign design and monitoring.	Friction points- links to outreach organizations, further information on harmful themes (digital literacy).	Exit Hate USA Former extremists as messengers, providing links to ways to leave extremism, Digital citizens campaign & Prebunking, enhancing resilience as opposed to reducing problematic behavior.	Focus on addiction treatment based approaches to extremist content.
Participant 3	Pressure to secure funding, celebrating success when there is none, lack of direct next steps, and follow-on action.	Integrated Mental Health Support.	Designed by civil society org, had a very real call-to-action for the target audience, involvement of civil society leaders in the design and delivery of the intervention.	Specific targeting of at-risk accounts.

Positive Interventions on Marketplace Platforms

Ellie Rogers, Swansea University

The GIFCT Blue Teaming Working Group (BTWG) identified marketplace platforms as a key space for terrorist and violent extremist (TVE) activity, but also a space where little is currently being done in terms of preventing and countering violent extremism (P/CVE) through potential positive interventions. As such, this section of the Playbook aims to provide guidance for tech companies, practitioners, policy makers, and academia by outlining how TVE actors are exploiting marketplace platforms and how these spaces can be utilized for positive interventions.

Introduction

The internet has grown in popularity and sophistication over the last decade, allowing individuals to communicate, share content, purchase and sell items, and build communities. TVE actors utilizing the internet are no exception to these developments, and have become active across various online spaces to spread propaganda,⁴⁷ recruit,⁴⁸ plan and coordinate attacks,⁴⁹ and fundraise.⁵⁰ One popular platform type that TVE actors have reportedly used is marketplace platforms,⁵¹ which are spaces for buyers and sellers of various products and services. These include sites such as Amazon Marketplace, eBay, Etsy, Pinterest, Depop, Shopify, Facebook Marketplace, Craigslist, Teespring, and Redbubble. Given the interconnected nature of the internet, marketplace platforms allow for products to be easily created and sold worldwide beyond traditional channels.⁵² As such, these platforms have been utilized to sell TVE products to generate funds, spread ideologies, and expand networks on a global scale.⁵³

Within P/CVE interventions, strictly prohibited TVE activity is typically addressed through methods of content removal. Borderline TVE activity, which is non-violent and does not overtly identify a proscribed

47 Anne Aly, Stuart Macdonald, Lee Jarvis & Thomas M. Chen, "Introduction to the Special Issue: Terrorist Online Propaganda and Radicalization," *Studies in Conflict & Terrorism* 40, no.1 (April 7, 2016), 1-9, <https://www.tandfonline.com/doi/abs/10.1080/1057610X.2016.1157402?journalCode=uter20>.

48 Gabriel Weimann, "Terror on Facebook, Twitter, And YouTube," *Brown Journal of World Affairs* 16, no.2 (2010), 45-54, <https://bjwa.brown.edu/16-2/terror-on-facebook-twitter-and-youtube/>.

49 Charlie Winter, Paul Neumann, Alexander Meleagrou-Hitchens, Magnus Ranstorp, Lorenzo Vidino and Johanna Fürst, "Online Extremism: Research Trends in Internet Activism, Radicalization, and Counter-Strategies," *International Journal of Conflict and Violence* 14 (March 3, 2021), 1-20, <https://www.ijcv.org/index.php/ijcv/article/view/3809>.

50 Ryan Scrivens, Tiana Gaudette, Maura Conway and Thomas J. Holt, "Right-Wing Extremists' Use of the Internet: Emerging Trends in the Empirical Literature," *Global Network on Extremism & Technology*, September 8, 2022, <https://gnet-research.org/2022/09/08/right-wing-extremists-use-of-the-internet-emerging-trends-in-the-empirical-literature/>.

51 Chris Joyner, "The Intersection of Online Retailers and Extremist Movements," *Government Technology* (blog), November 16, 2020, <https://www.govtech.com/security/the-intersection-of-online-retailers-and-extremist-movements.html>.

52 Ava Kofman, "Publishing Arm is a Haven for White Supremacists," *ProPublica* (blog), April 7, 2020, <https://www.propublica.org/article/the-hate-store-amazons-self-publishing-arm-is-a-haven-for-white-supremacists>.

53 Ben Stevens, "Amazon still selling extremist items like Nazi memorabilia despite pledge," *Retail Gazette* (blog), August 24, 2018, <https://www.retail-gazette.co.uk/blog/2018/08/amazon-still-selling-extremist-items-like-nazi-memorabilia-despite-pledge/>.

TVE group, is usually addressed through methods of content moderation (such as downranking, demonetizing, and limiting access) in addition to positive interventions (such as counterspeech, education, resilience building, and social cohesion) that are designed to address conditions conducive to radicalization, recruitment, or escalation.⁵⁴ However, P/CVE efforts currently focus on TVE use of social media platforms such as YouTube and Facebook, which creates a gap for TVE groups to remain high risk on marketplace platforms.⁵⁵

Based on interviews with marketplace platforms, focus group discussions with BTWG members, surveys disseminated to BTWG members, and a literature review, this section of the Playbook focuses on positive interventions on marketplace platforms. We begin by outlining marketplace platforms' general functions and surfaces and TVE exploitation of these spaces. We then discuss positive interventions that can be conducted on marketplace platforms. Finally, this output provides recommendations to improve future efforts of conducting positive interventions on marketplace platforms.

Marketplace Platforms and TVE Exploitation

Functions and Surfaces of Marketplace Platforms

Marketplace platforms allow for creating and selling a broad range of products, including custom, resale, and mass-produced items. Almost anyone can create and sell almost any item they choose on marketplace platforms. Popular items on online marketplaces are clothing, furniture and homeware, books, exercise equipment, electronics, and artwork. Unlike social media platforms, marketplace platforms are less communication-focused, and are more focused on allowing users to purchase items that align with their individual interests and daily lives. Marketplace platforms vary depending on the focus of what is being sold, which drives certain age demographics and geographic differences in terms of what platform is popular and where.

To sell an item on a marketplace platform, users create an account, then upload the product they are selling with images and a text description. To find a product to purchase, the most popular method is using the search function to seek out a specific item. Some platforms also allow users to shop by category, seller, or may have a home or listings page where users can see items posted by sellers they follow. Depending on the platform, users may be able to purchase items directly on the platform, but others have links to outsourced websites where users can purchase the item. Some (but not all) marketplace platforms also have user-to-user engagement and interaction features such as comments, a discussion section, a question-and-answer section, and direct messaging.

Marketplace platforms may also use recommendation algorithms, which utilize user information (such as browsing and purchasing histories) and information about the item (such as the category, seller, and

.....
 54 Erin Saltman and Micalie Hunt, "Borderline Content: Understanding the Gray Zone," Global Internet Forum to Counter Terrorism, <https://gifct.org/wp-content/uploads/2023/06/GIFCT-23WG-Borderline-1.1.pdf>.

55 Joyner, "The Intersection of Online Retailers."

keywords it is listed under) to link products that are commonly bought together or that the user may be interested in.⁵⁶ For example, many online book retailers use recommendation systems to direct customers who have shown interest in one book toward other similar books.⁵⁷ Sellers can also pay to have their products promoted on some platforms, so their item will be featured higher in search results, a listings page, or recommendations. On eBay (for example), sellers pay a percentage of an item's sale price, so the product is promoted higher in search listings and marked as sponsored.⁵⁸

TVE Exploitation of Marketplace Platforms

Due to pressure from governments, academia, and practitioners, marketplaces have introduced some restrictions to reduce activity associated with known terrorist organizations. There are not often specific TVE policies; instead, TVE products may fall afoul of legal or violence policies. For example, Amazon policies state that they block products that promote or glorify hatred, violence, racial, sexual, or religious intolerance, or promote organizations with such views.⁵⁹ Typically, harmful products are detected using automated machine learning, and human moderators review these products. However, it requires resources to train machine learning accurately, and human moderators often lack knowledge of TVE, and so require education. If TVE products are detected and deemed violative, they might be removed with no explanation, or users are provided with warnings and explanations as to why their products violate the platform's policies. This initial detection of TVE products can then cause a ripple effect for other similar products throughout the platform. For example, Etsy reportedly made efforts to remove QAnon merchandise in 2020,⁶⁰ and Amazon and eBay have reportedly made efforts to pull merchandise associated with The Proud Boys.⁶¹ There have also been instances of sellers being banned from platforms.⁶² However, these approaches leave gaps for borderline TVE products, which are still high risk on marketplace platforms.⁶³

The far right (referring largely to hate-based groups tied with white supremacy rhetoric) have the most recorded visibility on marketplace platforms, as right-wing extremist groups such as The Proud

.....
56 Kofman, "Publishing Arm."

57 Institute for Strategic Dialogue, Recommended Reading: Amazon's Algorithms, Conspiracy Theories and Extremist Literature, Institute for Strategic Dialogue, 2021, <https://www.isdglobal.org/wp-content/uploads/2021/04/Amazon-1.pdf>.

58 Adam Smith, "Qanon Products Including Neck Warmers and Earrings Available on eBay as Conspiracy Theory Spreads," The Independent, October 8, 2017, <https://www.independent.co.uk/tech/qanon-conspiracy-theory-ebay-products-trump-b886850.html>.

59 Kate Gibson, "Amazon, eBay Block Sale of Merchandise With 'Stand Back' and 'Stand By,'" CBS News, October 2, 2020, <https://www.cbsnews.com/news/stand-back-and-stand-by-proud-boys-merchandise-amazon/>.

60 Glossy Team, "Following an Etsy Ban, Qanon Merch is Still Readily Available on Online Marketplaces," ModernRetail, October 12, 2020, <https://www.modernretail.co/retailers/following-an-etsy-ban-qanon-merch-is-still-readily-available-on-online-marketplaces/>.

61 Glossy Team, "Following an Etsy Ban."

62 Kofman, "Publishing Arm."

63 Tom Keatinge, Florence Keen & Kayla Izenman, "Fundraising for Right-Wing Extremist Movements," Terrorist Financing 164, no.2 (May 30, 2019): 10-23, <https://doi.org/10.1080/03071847.2019.1621479>.

Boys, QAnon, Oath Keepers, III%ers,⁶⁴ and the Boogaloo Bois⁶⁵ have been found to sell products on marketplace platforms. There has also been documented jihadist extremist activity on certain platforms, but this appears to be less than the far right, potentially due to jihadist extremist groups typically being proscribed organizations⁶⁶ (meaning their products are likely banned outright on marketplace platforms). There is also far left anarchist activity on certain platforms.⁶⁷ However, TVE items are also commonly sold via third-party sellers, where the sale does not appear to benefit TVE groups directly and the seller also offers a variety of non-TVE merchandise as well.⁶⁸

Interviews with marketplace platforms highlighted that custom-made items were a significant problem, as creating and selling products with an image, symbol, or slogan associated with TVE is relatively simple. For example, swastika coins, German tank jackets with the Nazi swastika and SS symbols, and stamps commemorating Hitler were found to be available on Amazon⁶⁹ and Facebook Marketplace.⁷⁰ These products may also be amplified by algorithms, increasing their visibility on platforms. One example stems from eBay, where there have been automatic suggestions for QAnon products.⁷¹ It is also hard to detect these items for several reasons. Symbols can be subjective and highly dependent on context; for example, mythic Norse symbology has been found to be exploited by the extreme right.⁷² Custom items are also often listed under coded or shorthand keywords⁷³ (for example, patches and t-shirts associated with the Boogaloo Bois were listed on Amazon under less obvious keywords).⁷⁴ If explicit search terms such as white power, swastika, and skinhead are used on marketplace platforms, no problematic results are often found.⁷⁵ However, this may not be the case for all groups and platforms.

Books published and sold by TVE actors (which often contain hate-based and extreme themes) are

64 ADL, "Tech Companies Still Supporting Hate, ADL (blog), January 7, 2021, <https://www.adl.org/resources/blog/tech-companies-still-supporting-hate><https://www.adl.org/resources/blog/tech-companies-still-supporting-hate>.

65 Michael Waters, "How Amazon's Algorithm Allows Extremist Merchandise to Spread," ModernRetail, January 15, 2021, <https://www.modernretail.co/retailers/how-amazons-algorithm-allows-extremist-merchandise-to-spread/>.

66 Felix Pope, "Exposed: Jihadi Propaganda on Sale on Amazon," The JC, October 20, 2022, <https://www.thejc.com/news/news/exposed-jihadi-propaganda-on-sale-on-amazon-lMtcDsQ0oXcProaQCcrwiX8>.

67 Joyner, "The Intersection of Online Retailers."

68 ADL, "Amazon: Hate for Sale," ADL (blog), July 29, 2019, <https://www.adl.org/resources/blog/amazon-hate-sale>.

69 Stevens, "Amazon Still Selling."

70 Sarah Chemla, "Anti-Defamation Commission Condemns Facebook for selling Nazi-like Items," The Jerusalem Post, August 14, 2020, <https://www.jpost.com/diaspora/antisemitism/anti-defamation-commission-condemns-facebook-for-selling-nazi-like-items-638659>.

71 Smith, "Qanon Products."

72 Tom Birkett, "Far-right Extremists Keep Co-Opting Norse Symbolism - Here's Why," The Conversation (blog), June 16, 2022, <https://theconversation.com/far-right-extremists-keep-co-opting-norse-symbolism-heres-why-183749>.

73 Tim Squirrell and Clara Martiny, Profiting From Hate: Extremist Merchandise on Redbubble, Etsy, Teespring, Teerepublic and Zazzle, Institute for Strategic Dialogue, 2022, <https://www.isdglobal.org/wp-content/uploads/2022/12/Profiting-from-Hate-Extremist-Merchandise-on-Redbubble-Etsy-Teespring-Teerepublic-and-Zazzle.pdf>.

74 Katie Canales, "Weeks After Saying it Removed Boogaloo Products, Amazon is Still Listing Patches and T-Shirts Promoting the Far-Right, Extremist Group," Insider (blog), August 13, 2020, <https://www.businessinsider.com/amazon-sellers-listing-boogaloo-related-products-2020-8?r=US&IR=T>.

75 ADL, "Amazon: Hate for Sale."

also commonly sold on marketplace platforms, as they are not explicitly illegal items. In 2019, a QAnon book was among the top 50 best sellers on Amazon.⁷⁶ There are also concerns surrounding algorithmic optimization of these items: The Turner Diaries, which is a white supremacist book, has been proactively recommended by Amazon's algorithm.⁷⁷ This amplification may be consciously exploited by TVE actors. Sellers can make their books free so users can easily read them and leave positive reviews, resulting in higher-ranking placement in search results.⁷⁸ "Category squatting" has been utilized as well, where sellers classify books in obscure, low-traffic categories to improve their rankings.⁷⁹ As well as increasing the visibility of TVE books, this optimization can be problematic, as research has linked recommendations of far right texts with the creation of extremist rabbit holes.⁸⁰

There have also been instances of marketplace platforms being associated with weapon building and TVE. For example, in 2017, Amazon's algorithm grouped together materials that can be used to create explosives, suggesting these items were commonly purchased together.⁸¹ In 2021, Facebook advertised military equipment alongside content promoting election misinformation.⁸² As these items are not explicitly illegal or often sold by TVE-affiliated accounts, determining intent and regulating this activity can be difficult.

During interviews with marketplace platforms, it was noted that one of the most challenging problems surrounding TVE products is that individuals may not understand why a product is harmful due to a lack of knowledge regarding TVE iconography. For example, users may not be aware that a t-shirt listed as 'vintage' displays a neo-Nazi symbol. A lack of education may also mean that individuals sell products on behalf of groups without knowing their background or core values. This is particularly seen within mainstream popular culture, as influencers and users alike may not be aware of a product's association with TVE. QAnon merchandise (for example) is often linked to female lifestyle and wellness influencers.⁸³ This association with popular culture is exacerbated by the easy creation of custom products, as slogans, symbols, and memes can be rapidly incorporated into products. As such, TVE products on marketplace platforms may differ over time, consistent with popular trends.

Given the examples of exploitation discussed above, there is considerable need for intervention assistance to help marketplace platforms.

.....
76 Waters, "How Amazon's Algorithm."

77 ADL, "Amazon: Hate for Sale."

78 Kofman, "Publishing Arm."

79 Kofman, "Publishing Arm."

80 Jessica Guynn, "Browsing Recommended Books on Amazon Can Lead to Extremist Rabbit Hole," Tech Explore, April 29, 2021, <https://techxplore.com/news/2021-04-browsing-amazon-extremist-rabbit-hole.html>.

81 Thomas McMullan, "Amazon 'Reviewing' Algorithms That Promoted Bomb Materials," Alphi, September 19, 2017, <https://www.alphi.com/politics/1007077/amazon-reviewing-algorithms-that-promoted-bomb-materials/>.

82 Ryan Mac and Craig Silverman, "Facebook Has Been Showing Military Gear Ads Next To Insurrection Posts," BuzzFeed News, January 14, 2021, <https://www.buzzfeednews.com/article/ryanmac/facebook-profits-military-gear-ads-capitol-riot>.

83 Glossy Team, "Following an Etsy Ban."

Positive Interventions and Marketplace Platforms

As part of wider counter terrorism initiatives to address TVE use of the internet and the increasing global threat they pose, P/CVE interventions are conducted online. Positive interventions aim to deter individuals from exploiting online platforms, disengage users participating in TVE, prevent them from becoming involved in TVE, build resilience to radicalization, and empower the general public to become involved in P/CVE efforts.⁸⁴

Upstream Interventions

The overarching suggestion by the BTWG was improved detection and moderation of TVE products on marketplace platforms. This can be done by improving automated efforts by marketplace platforms and increasing human moderator efforts, but BTWG members also suggested to involve users in flagging TVE products. This may require incentivizing users to become involved in addressing TVE activity through free credits to use on a platform, the platform donating to a relevant cause for each product flagged (see Case Study 1), or for users' products to be promoted on platforms if they actively identify TVE products. However, it is important to accompany efforts to incentivize users with education and increased transparency to improve the accuracy of flagging. This could be delivered as a short education module that users must click through before joining the platform. The module could inform users on how to utilize platform tools to flag and report products (see Case Study 2), what TVE is, what products are associated with TVE, why the sale and purchasing of these products is harmful, and what signs and keywords to look out for.

Case Study 1: The "Donate the Hate"⁸⁵ program informed users how to flag hateful and TVE comments and provided rewards in the form of donations to charity. Over €10,000 was raised to fund further non-governmental (NGO) disengagement services. Something similar could be developed on marketplace platforms to turn the presence of TVE products and the prevention of their sale into something positive.

Case Study 2: The Anti-Defamation League (ADL) created the Cyber-Safety Action Guide,⁸⁶ which outlines how to identify and report hate online. This was developed for broader online hate and harassment, but something similar could be replicated in the TVE space.

The BTWG also suggested that platforms could complete training and take steps to become a certified

.....

⁸⁴ Global Internet Forum to Counter Terrorism, Content-Sharing Algorithms, Processes, and Positive Interventions Working Group- Part 2: Positive Interventions, Global Internet Forum to Counter Terrorism, 2021, <https://gifct.org/wp-content/uploads/2021/07/GIFCT-CAPI2-2021.pdf>.

⁸⁵ To read more about the 'Donate the Hate' campaign, see https://www.hasshilft.de/index_en.html.

⁸⁶ To read more about the Cyber-Safety Action Guide, see <https://www.adl.org/cyber-safety-action-guide>.

ethical platform (see Case Study 3). Platforms could then involve users in this goal by remaining ethical sellers, which can aid in promoting community cohesion across the platform.

Case Study 3: B Corporation⁸⁷ offers information detailing how platforms can gain ethical certification and have a positive societal and environmental impact. This certification focuses primarily on environmental factors, but something similar could be replicated for TVE, where platforms undergo and adhere to P/CVE training.

BTWG members also highlighted that the context and seller of the product are often more important than the product itself, as innocent products can be sold in harmful ways. Consequently, educational modules could explain how seemingly harmless products can be sold by users aiming to bring someone into a TVE community or lead them down a harmful pathway.

The BTWG suggested that humor could be used to make information more engaging and light-hearted while proactively setting norms. However, emotion-driven content remains very context-dependent and should be used with a deep knowledge of the target audience to reflect social sensitivities. Alternatively, marketplace platforms suggested that referencing how the sale and purchasing of TVE products can negatively impact or harm others would likely induce more caring responses than a broad, impersonal message. Partnerships with subject matter experts may be beneficial for designing messages, and continual monitoring and evaluation need to take place to make adjustments if interventions are ineffective.

This combination of incentivizing and educating users on how to be involved in identifying and flagging TVE products may be an effective intervention, regardless of whether it increases the flagging of TVE products, as it could reduce the sale and purchase of TVE products by providing education on why specific phrases and symbols are associated with TVE. This could also have a domino effect on other platforms by users carrying this knowledge forward and educating others (see Case Study 4).

Case Study 4: Australia's Community Action for Preventing Extremism's (CAPE) "Exit White Power" project⁸⁸ was set up in 2012 by All Together Now. The website provides educational information about white supremacy in Australia and enacted in-depth thematic discussions.⁸⁹ Marketplace platforms could deliver something similar to users before they join platforms.

87 To read more about the B Corporation, see <https://www.bcorporation.net/en-us/certification/>.

88 To read more about the 'Exit White Power' project, see <https://alltogethernow.org.au/our-work/far-right-extremism/community-action-for-preventing-extremism-cape/>.

89 William Alcorn, Australian Radical Right Narratives And Counter-Narratives In An Age Of Terrorism, Hedayah, 2021, https://hedayah.com/app/uploads/2021/09/002_Final_CARR-Hedayah_Australia-Country-Report_2021JAN26.pdf.

However, it is important that all approaches be enacted sensitively by marketplace platforms with effective measurement and evaluation to ensure their quality and be kept up-to-date with changing trends in TVE activity. For example, the BTWG suggested having more LGBTQ+ inclusivity within information during Pride month to account for potential increases in hateful products.

Another upstream approach that the BTWG suggested was P/CVE practitioners utilizing marketplace platforms to sell their own products to raise funds and awareness of P/CVE work (see Case Study 5).

Case Study 5: The German organization “Loud Against Nazis”⁹⁰ sell coffee and merchandise to fundraise (for example, they sell t-shirts⁹¹ saying “Good morning love, goodbye hate”). Similar organizations could sell their products on marketplace platforms.

As is the case with other approaches, continual monitoring and evaluation is necessary to assess the effectiveness of this strategy.

Downstream Interventions

The BTWG suggested explaining why a product has been flagged or removed for being associated with TVE (an approach already utilized on some platforms). For products flagged as being associated with TVE but which can remain on platforms, explanations could appear next to the product. For products removed for being associated with TVE, sellers could be required to read the information before regaining access as a seller. Moreover, to cater to the purpose of marketplace platforms, the BTWG suggested offering alternative products to flagged or removed products, or providing sellers with links to support services if their products are removed or flagged for being associated with TVE. However, all approaches need adequate measurement and evaluation, with the flexibility to evolve and change if the method is ineffective.

Another suggestion by platforms and the BTWG was to utilize search redirect. Pioneered by Jigsaw and Moonshot⁹² in 2016, the Redirect Method uses platform search engines and ad tech algorithms to surface alternative content when a user searches for predetermined keywords. This method is already employed on social media platforms for TVE and other harmful content (see Case Study 6), and by certain marketplace platforms to redirect users searching for items related to suicide and self-harm. Search redirect could be adapted to align with marketplace platform functions. For example, if a user searches for a white supremacist book containing hate speech or written by a TVE actor, they could be redirected to a non-harmful educational book on a similar topic. To offer individuals access to safe communities (as buying behavior often reflects the community users are a part of), links to an outsource group relevant

90 To read more about ‘Loud Against Nazis’ organization, see <https://www.lautgegennazis.de/>.

91 To read more about ‘Loud Against Nazis’ merchandise, see <https://krasserstoff.com/lautgegennazis>.

92 Moonshot, The Redirect Method, Moonshot, <https://moonshotteam.com/the-redirect-method/>.

to the TVE item being searched for could be included (see Case Study 7).

Case Study 6: “Stop It Now!”⁹³ created a campaign where users attempting to search for illegal child sexual abuse (CSA) content or visit websites on a banned list were redirected to the Get Help website,⁹⁴ which offers resources to address this behavior and stop offending. The pilot was effective in driving users toward the helpline and self-help resources. There were self-reported changes of a better understanding of the laws and impact of offending, as well as reduced offending behavior. This example could act as an evidence base for informing TVE redirect campaigns on marketplace platforms.

Case Study 7: Facebook’s redirect campaign⁹⁵ aims to redirect those searching for TVE-related content towards outreach groups and educational resources. The pilot campaign found that 25 individuals who were seeking to engage with white supremacist content in the US engaged with Life After Hate⁹⁶ instead.⁹⁷ Something similar could be adapted for marketplace platforms.

Some evidence suggests redirection is effective on marketplace platforms, as one platform reported that individuals who were shown educational messages after searching for harmful products did not look further for harmful products. However, as with any P/CVE campaign, measurement and evaluation is essential to determine whether the approach is effective. Platforms had concerns about this approach deterring individuals from the platform and felt that redirection should not interrupt user experience too much.

It can also be challenging to determine which keywords should trigger redirection, due to TVE items commonly being sold under less obvious keywords, and to determine the products and outreach groups users would be redirected to. Previous P/CVE work has utilized subject matter experts to devise a list of TVE keywords, which could be a beneficial partnership for marketplace platforms. More research and continual monitoring of groups, popular language, and trends is needed to update keywords and what users are directed toward.

Buying and searching habits alone cannot tell us much about a user’s intentions, as there are many

93 To read more about ‘Stop It Now!’, see <https://www.stopitnow.org.uk/home/media-centre/news/deterrence-campaign-faithfull-paper/>.

94 To see the ‘Get Help’ page for ‘Stop It Now!’, see https://www.stopitnow.org.uk/concerned-about-your-own-thoughts-or-behaviour/concerned-about-use-of-the-internet/self-help/understanding-the-behaviour/images-are-children/?_gl=1*5eecp3*_up*MQ..*_ga*MTQ2Mzc5NjlyOC4xNjkxNDI-4ODY5*_ga_STZD47XNW7*MTY5MTQyODg2OC4xLjEuMTY5MTQyODg2OC4wLjAuMA.

95 To read more about Facebook’s Redirect campaign, see <https://counterspeech.fb.com/en/initiatives/redirect/>.

96 To read more about Life After Hate, see <https://www.lifeafterhate.org/>.

97 Moonshot, From Passive Search to Active Conversation: An Evaluation of the Facebook Redirect Programme, Moonshot, 2020, https://counterspeech.fb.com/en/wp-content/uploads/sites/2/2020/11/Facebook-Redirect-Evaluation_Final-Report_Moonshot-1.pdf.

different reasons someone may purchase a product (for example, academics often purchase books associated with TVE for their research). As such, another suggested method by the BTWG for identifying target audiences is to identify TVE influencers and accounts to determine which users have interacted with them and their products. Depending on the platform's surfaces, alternative products and links to outreach groups could be offered to these users through ad targeting, recommendation algorithms, or one-to-one interventions (see Case Study 8). Partnerships with NGOs may be beneficial for one-to-one interventions, which require considerable resources.

Case Study 8: The Institute for Strategic Dialogue's (ISD) Counter-Conversations program⁹⁸ identified individuals showing signs of radicalization on social media and directly engaged with them through private and personalized messages. The program saw that one in ten individuals who were contacted suggested it had a positive impact. A similar approach could be conducted on marketplace platforms with engagement surfaces.

Delivering Positive Interventions

Marketplace platforms interviewed for this chapter noted that private companies are not the best "authentic voice" for delivering positive intervention messaging. Research has also shown that using "credible messengers"—sometimes in the form of influencers or popular voices in the online space (see Case Study 9)—could help give meaning and ensure a greater impact for positive interventions.⁹⁹ This would depend on the surface and intervention is deployed within marketplaces and in close partnership between said influencers and the platform. In all cases, interventions should have a measurement and evaluation approach decided upon before the launch of a campaign or intervention.

Case Study 9: Search for Common Ground (SFCG) Indonesia developed Generating Indonesian Resilience and Leadership Skills (GIRLS),¹⁰⁰ which focused on creating opportunities for young women who may be targeted by TVE groups.¹⁰¹ SFCG collaborated with young micro-influencers, who were encouraged to incorporate messages of tolerance into their regular content. This was deemed a beneficial partnership, as it utilized a credible messenger, which is a crucial factor in the efficacy of campaigns. Similar partnerships could be used by marketplace platforms.

98 To read more about ISDs Counter-Conversations programme, see https://www.isdglobal.org/wp-content/uploads/2018/03/Counter-Conversations_FINAL.pdf.

99 Munir Zamir, Active Strategic Communications: Measuring Impact and Audience Engagement, GIFCT Positive Interventions and Strategic Communications Working Group, 2022, <https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-PI-Impact-1.1.pdf>.

100 To read more about GIRLS campaign, see https://www.rsis.edu.sg/wp-content/uploads/2020/09/GNET-CENS-Workshop_Online-Agitators-Extremists-and-Counter-Messaging-in-Indonesia.pdf.

101 S. Rajaratnam School of International Studies, Online Agitators, Extremists and Counter-Messaging in Indonesia, S. Rajaratnam School of International Studies, 2020, https://www.rsis.edu.sg/wp-content/uploads/2020/09/GNET-CENS-Workshop_Online-Agitators-Extremists-and-Counter-Messaging-in-Indonesia.pdf.

Conclusion and Recommendations

In conclusion, TVE actors use marketplace platforms to sell and purchase a range of products in order to fundraise and spread their ideologies. What follows highlights a range of positive upstream and downstream interventions that could be conducted on marketplace platforms. For any successful approach, the intervention, messenger, and platform must be aligned, which requires collaboration from the different sectors.¹⁰² As such, the BTWG has put forward recommendations that will aid in gaining a better understanding of TVE activity on marketplace platforms and to conduct P/CVE positive interventions more effectively from a multi-stakeholder approach.

Recommendations for Marketplace Platforms

Positive Interventions: The first upstream intervention the BTWG suggests is involving users in flagging TVE products by incentivizing users with something positive and providing education modules for users to read as part of joining the platform. The second upstream intervention that the BTWG suggests is for P/CVE organizations to sell their products on marketplace platforms to fundraise and raise awareness of their work. The first downstream intervention the BTWG recommends is to provide explanations, alternative products, or links to support services when a product is flagged or removed for being associated with TVE. The second downstream intervention is to redirect users searching for TVE products towards alternative products and support services. The third downstream intervention the BTWG suggests is to surface alternative products or support services through ad targeting, recommendations, or one-to-one interventions with users who interact with or purchase products from identified TVE accounts. The BTWG also notes the potential benefits of working with influencers to deliver positive interventions. Finally, the overarching suggestion from the BTWG is that all approaches are carefully monitored and evaluated to make adjustments if they are ineffective (see Appendix for more information on platform surfaces and recommended positive interventions).

Transparency: An important recommendation for marketplace platforms is increased transparency surrounding the scale of TVE activity and detection and moderation approaches. For users, this will mean having more information on how their products are being dealt with and why this is the case. For academics and practitioners, it will allow for more effective research on TVE use of marketplace platforms and how to develop appropriate responses. Furthermore, it will provide academics, practitioners, and NGOs with more information on measuring, monitoring, and evaluating positive interventions, which is key for adhering to the “do no harm” principle and achieving effective interventions. This will also better inform policy and legislation by providing policy makers with a clearer understanding of the threat of TVE activity on marketplace platforms.

Detection and moderation: Platforms should seek to improve automated detection and flagging of violative and non-violative TVE. This should be accompanied by improved training and education of

.....
 102 Zamir, “Active Strategic Communications.”

human moderators to assess flagged products and determine what action should be taken. As well as ensuring the most extreme products are removed, this is necessary to address borderline TVE products and conduct positive interventions, as identifying and flagging products is often the first step for target audience identification. Partnerships with academia and NGOs could be used to ensure moderators are trained in TVE detection and moderation.

Consistency: While a one-size-fits-all approach is not possible, there should be some consistency between marketplace platforms, as removing and addressing TVE activity on one platform could make it more likely that the problem is merely displaced onto another platform. Achieving some consistency could be aided by increased government and policy maker focus on marketplace platforms within legislation, as international alignment would ensure marketplace platforms are held accountable for TVE products they are profiting from and are building robust and transparent processes to detect, remove, and deter TVE exploitation of their platforms.

Partnerships: It is important for platforms to form partnerships with academics, practitioners, and NGOs to gain the expertise and resources to understand TVE activity on marketplace platforms and ethically and effectively carry out positive interventions. It is also essential for platforms to form partnerships with outreach organizations that users are linked to within interventions to ensure users have access to any support they may require.

Recommendations for Civil Society (NGOs, P/CVE Practitioners, and Academia)

Research on TVE: Further longitudinal research into TVE activity on marketplace platforms would be beneficial, aided by increased transparency from platforms. These will give further insight into the range and type of TVE activity and products being purchased and sold on marketplace platforms over time, and how these transform in line with popular culture and alterations in platform operations. This research would be beneficial to identify the problem and target audiences effectively, as well as design and deliver positive interventions relevant to the platform and the types of TVE products being engaged with.

Measurement, monitoring, and evaluation: Whatever interventions are used, it is important that monitoring and evaluation are conducted to ensure they are effective and ethical and have a positive impact without counterproductive or negative consequences. This includes further research on the long-term impact of these interventions, which can be aided by increased transparency from marketplace platforms on what data and metrics can be used. From measurement and evaluation findings, frameworks can be developed for other platforms to draw upon when conducting their own positive interventions.

Partnerships: As well as forming partnerships with marketplace platforms, collaboration among academics, practitioners, and NGOs would be beneficial to share knowledge and expertise. This will ensure multi-disciplinary knowledge on TVE use of marketplace platforms when conducting and evaluating positive interventions in this space. Partnerships with outreach organizations will also allow

for multi-stakeholder input into how to cater interventions specifically to marketplace platforms and the types of TVE activity on each platform.

Recommendations for Government and Policy Makers

TVE and legislation: As is the case for social media platforms, it is vital that governments and policy makers acknowledge TVE activity on marketplace platforms within legislation, particularly for clearly violative products. Such legislation will require marketplace platforms to monitor and detect products on their platforms more closely to disrupt and remove illegal TVE products. This may also require marketplace platforms to develop specific TVE policies to address this content appropriately. Partnerships with academia and NGOs could help develop these policies.

Borderline TVE and legislation: The majority of TVE activity on marketplace platforms is classified as borderline or non-violent, so more resources and attention need to be dedicated to this activity. There have been recent developments for addressing so-called legal but harmful content on social media platforms under the United Kingdom's Online Safety Bill, but this leaves gaps surrounding borderline TVE activity on marketplace platforms. There should be alignment on what illegal terrorist products should not be sold on platforms. Then, platforms and academia should gather evidence on the scale and nature of borderline activity to inform a response. This process can be informed by how other online harms are dealt with, such as borderline CSA material and suicide and self-harm content.

Appendix: Marketplace Platforms and P/CVE Positive Interventions Matrix

Surfaces	Positive Intervention					
	Search Redirect	Education & incentives for user-led flagging ¹⁰³	P/CVE organizations selling products	Education & alternatives for flagged and removed TVE products ¹⁰⁴	One-to-one interventions ¹⁰⁵	Ads promoting alternative products ¹⁰⁶
Direct messaging ¹⁰⁷				X	X	
Home / listings page ¹⁰⁸			X	X		X
Search function ¹⁰⁹	X		X			
Joining page ¹¹⁰		X				
Product page ¹¹¹				X		
Recommended products ¹¹²			X			X

.....

103 Upstream education shown to users before they join the platform.

104 Downstream interventions shown after flagging or action are taken on a TVE product.

105 For users interacting with or purchasing products from an identified TVE account or influencer.

106 For users interacting with or purchasing products from an identified TVE account or influencer.

107 Refers to comments or discussion pages for products, or platforms that have direct messaging options once interest has been placed in a product.

108 Where users can browse for products, see promoted or sponsored products, and advertisements.

109 Where users can search for specific products.

110 When users join a platform, make an account, and read the terms of service and community guidelines of the platform.

111 Refers to the page of a specific product that users can click on if they wish to view the description and photographs of the product.

112 Recommended products are typically shown as a sidebar or in a section on the home or listings page, or on the page for a specific product.

Strategizing Online Interventions In Global Contexts

Prof. Fredrick Ogenga, Center for Media, Democracy, Peace & Security, Rongo University;
The Peacemaker Corps Foundation Kenya

Introduction and Overview

This output considers key questions and considerations for practitioners and technology companies as they strategize online interventions. These considerations have often been taken for granted, leading to fragmented approaches to online interventions on preventing and countering violent extremism (PVE/CVE), only to be further exacerbated by tech challenges and gaps between the global North and South. Tech companies, civil society, and arguably state agencies need to consider geographical, cultural, and demographic variables when designing and implementing positive intervention strategies. These strategies must be anchored by key issues that act as guardrails for monitoring success. For example, positive interventions adopted online ought to consider critical contextual or background variables such as platform selection, message formulation, and audience alignment. They should also consider digital ethnography, radicalization pathways, background research on specific harm vectors, and alternative platforms.¹¹³

Regarding platform choice, messaging, and audience alignment, there is a pressing need to look at particular platforms that are aligned with distinct audiences, groups, identities, behaviors, and extremism vectors. It is vital to assess interventions tailored for group and individual levels¹¹⁴ and to assess the degree of alignment across message, messenger, and platform for different platforms. Digital ethnography and radicalization pathways should be analyzed to broaden conceptualizations of at-risk audiences for violent extremism. Questions arise regarding the feasibility of reverse-engineering the recruitment process to facilitate disengagement work.¹¹⁵ Whether ethnographic research can be made actionable for disengagement remains undetermined.¹¹⁶

Due to the possibility of cross-platform interactions like link and content-sharing, people and content migration across platforms, and comments serving as discussion starters, research on specific harm vectors becomes imperative. There is the possibility of using alternative platforms for redeploying messaging from existing platforms and using narrative strategies such as storytelling to counteract extremist narratives.¹¹⁷ Ideally, these technical, geographical, cultural, and demographic considerations put the user at the center of this output. There is a critical need to consider audiences and users as integral

113 See GIFCT (2022). Alternative Platforms for Positive Intervention 2022-2023. Blue Teaming Working Group: GIFCT; Responses from participants. Unpublished.

114 Alternative Platforms, BTWG.

115 Alternative Platforms, BTWG.

116 Alternative Platforms, BTWG.

117 Alternative Platforms, BTWG.

components in the research and development of platform functionalities. Their implications for positive interventions – especially within the realms of tech and society and especially in the discourse around PVE/CVE online – are crucial in the context of fashioning a critical theory of technology.¹¹⁸

Critical Theory of Technology

As an alternative approach the dominant approaches to technology—found in theories such as those of technological determinism, which argues that technology determines our destiny – this theory is used in this output to emphasize contextual aspects of technology ignored by the dominant view. Technology is not only the rational control of nature; both its development and impact are intrinsically social. The authoritarian social hierarchy of technology should not be viewed as a contingent dimension of technical progress but as a technical necessity for the preservation of power modalities that ought to be resisted. Terrorists use technology to propagate fear, hate, and death to advance power interests online. Users must be educated, incentivized, trained, and made aware of how technology supports those interests and what they can do to avoid vulnerability. This should be a comprehensive societal conversation that brings together academics, policy makers, civil society, tech companies, and, most importantly, users. Considerations should be made in terms of the type of education needed for the different levels of risk, e.g., those who might be vulnerable to extremist content but are not yet engaged (early-stage risk) versus those who are actively seeking out terrorist content (later-stage risk). Being able to categorize user behavior into different risk levels could support a more nuanced and targeted approach.

Methodology

While other deliverables for the Blue Teaming Working Group looked at the specific viability of PVE/CVE intervention strategies across various platforms and online surfaces, this output sought to identify key questions and considerations that practitioners and tech companies should review when formulating online intervention strategies. It emphasized the proactive assessment and addressing of geographic, cultural, and demographic factors to ensure that interventions yield positive outcomes without inadvertently causing harm. The output used monthly Working Group Virtual Sessions as a data-gathering technique to prime questions and develop a perspective for the paper. Data was collected in real time from a series of monthly meetings with experts, including tech company representatives and other stakeholders. A methodological weakness arose out of the use of Chatham House rules, which limited the extent of the application of data from respondents in terms of identity disclosure. Nevertheless, the real-time setting fostered candid and authentic responses that shed light on gaps and opportunities for using different social media surfaces/platforms for positive interventions across geographic spaces, demographics, ecological contexts/localities, ethnicities, and cultures.

.....
118 Andrew Feenberg, "Democratic Rationalization: Technology, Power and Freedom," in *Philosophy of Technology*, eds. Robert Scharff and Val Dusek, 2nd ed. (Oxford: Wiley Blackwell, 2014), 706–726.

Disjointed Interventions? North-South Technological Divisions

It appears the main gap stems from the fragmented nature of current interventions across different geographic spaces and the limited comprehension of how audiences in the global South apply technology. Technology (especially social media) is exploited differently in various geographical contexts.¹¹⁹ Given that most tech emerges from the North, the moral implications of technology on this audience should be based on how they choose to apply it. Considering these exposed numerous challenges, ranging from the policy environment and cultural barriers to access to technology (especially gaming platforms) and tech-literacy.¹²⁰ Despite the fact that virtually all these challenges are more prevalent in the global South, there is limited evidence of how tech companies (predominantly domiciled in the North) and other stakeholders are working to address them at the global level. This results in the prioritization of useful positive interventions online for PVE/CVE to spaces where these challenges do not seemingly exist.

Collaborative Collective Model

However, there are grounds for optimism, judging from the degree of congruence seen in how most surfaces have interventions that seek to nurture the positive application of technology. These can be used to build up a collaborative, collective model across geographies, demographics, cultures, ethnicities, localities, and contexts (platform sharing of messaging/interventions), with the major challenge simply being that some of these surfaces are competitors. Nevertheless, if the common denominator behind positive interventions is to advance “tech for good”¹²¹ around PVE/CVE, then an opportunity exists to redesign and define mechanisms that can be used by tech surfaces to break the geographical, demographic, ethnic, cultural, and contextual limits and barriers.¹²²

The study underscores the limited popularity of certain platforms in regions like Africa, such as the discovery that most users of some of these surfaces (especially gaming platforms) are not based in Africa. Variables like culture play a large role in whether or not gaming platforms (for instance) remain unpopular, and underscores the significance of audience exploitation when examining online violent extremism and the strategies to counteract it through positive interventions.¹²³ At the same time, other platforms (e.g., marketplaces) are also utilized differently in the African context, dictating that intervention approaches cannot be universalized. Different consumer habits regarding shopping online are shaped

.....
119 See Martin Ndlela and Abraham Mulwo, “Social Media, Youth and Everyday Life in Kenya,” *Journal of African Media Studies* 9, no. 2 (2017): 277–290, <https://www.scholars.northwestern.edu/en/publications/social-media-youth-and-everyday-life-in-kenya>.

120 See Kate Cox, William Marcellino, Jacopo Bellasio, Antonia Ward, Katerina Galai, Arya Meranto, and Giacomo Paoli, “Social Media in Africa: A Double-Edged Sword for Security and Development,” United Nations Development Programme (UNDP) Regional Centre for Africa (November 2018), https://www.rand.org/pubs/external_publications/EP67728.html.

121 See Alison Powell, A., Funda Ustek-Spilda, Sebastian Lehuedé, and Irina Shklovski, “Addressing Ethical Gaps in ‘Technology for Good’: Foregrounding Care and Capabilities,” *Big Data & Society* 9, no. 2 (July–December, 2022), <https://journals.sagepub.com/doi/epub/10.1177/20539517221113774>.

122 See Lisa Schirch, “A Peacebuilding Approach to Addressing Social Media Threats,” in *Social Media Impacts on Conflict and Democracy: The Tonic Shift*, ed. Lisa Schirch (New York: Routledge, 2021), 216–234.

123 Cox et al., “Social Media in Africa.”

by their locality, income status, social norms, and existing legislation that determines public access to technology. This points to a critical consideration: involving users in defining their role within positive interventions may inadvertently overlook potential contributions from other geographic spheres.

Platform Exploitation and Positive Interventions

These outcomes underscore the impracticality of universally applying the same approach toward preventing and countering violent extremism in unexploited gaming platforms, marketplaces, and nouveau platforms. Consideration should be given to how to classify immediate versus emerging threats for different locations. For example, gaming platforms might be an emerging threat for right-wing extremists, but are gaming platforms a similar threat for Islamic extremists based in sub-Saharan Africa? Questions like how likely a violent extremist organization (VEO) is to take up a certain technology or how North/South partners could work together to use one area's immediate threats to help prepare for another area's emerging threats are not currently being asked. There is virtually no research looking at Africa-based jihadists on gaming platforms. Nevertheless, just because right-wing extremists are loud and obvious in these spaces does not mean there are no jihadis on these platforms.

This should be an important question for practitioners and tech companies. It is a question that equally invites context-specific baseline studies that would allow for the introduction and orientation of potential users on those platforms and argue for their potential future contribution to PVE/VE at a universal scale. For example, if you examine one of al-Shabaab's most publicized attacks – the Westgate attack¹²⁴ on September 21, 2013 – there was notably little engagement between al-Shabaab's Twitter accounts and individual Twitter users. Only two percent of the 556 tweets issued from al-Shabaab's eight Twitter accounts during the attack involved communicating directly with individual Twitter users, and none of those tweets led to a discussion or platform exploitation.¹²⁵ One explanation for this lack of engagement with individual Twitter users is a reluctance to give others control of the narrative, thus diverting attention from al-Shabaab's communications at a time of high publicity. Careful attention should be paid to how users exploit technology and when and how extremist groups choose to engage users with extremist content online. In the aforementioned instance, al-Shabaab flooded Twitter with horrific imagery and real-time content about the attack, so their goal might have been to increase terror and not necessarily to recruit or engage with new users. One day, jihadis may also livestream their attacks on Twitch, DLive, etc., for the same reasons; whether or not they will is another potential area for research.

Target Audience/Users, Contextual Nuances, and Grassroots Collaboration

What is clear, though, is that government and stakeholder regulation, along with non-regulatory lobbying, can also serve as factors in how platforms prioritize where to target. In fact, many factors go into how platforms prioritize where to target, including the exploitation of technology by audiences

.....
 124 See John Campbell, "Justice, Terrorism and Nairobi's Westgate Mall," Council of Foreign Relations, October 9, 2020, <https://www.cfr.org/blog/justice-terrorism-and-nairobis-westgate-mall>.

125 Campbell, "Justice, Terrorism, and Nairobi." Five of the group's Twitter accounts responded to specific journalists and other individuals.

(and vice versa) and available partnerships on the ground to make positive interventions work. Other factors include potential conflict fault lines like high-risk elections in some countries and geographic regions and highly sensitive moments to which platforms must allocate limited resources.¹²⁶ Another factor is generating metrics for understanding and measuring prevention efforts. It is therefore always a challenge to figure out what the right areas to target are.

Some platforms (such as video messaging platforms like TikTok or Facebook Reels) may choose regions based on internal metrics that identify the worst potential search terms and where they can establish partnerships most easily, which often leaves out spaces in the global South.¹²⁷ Setting up partnerships on the ground with NGOs is critical to ensuring the successful launch of campaigns. For example, some platforms have developed stronger partnerships in the Asian and Pacific markets than Africa and other regions due to the high volume of available grassroots partners. The ease of access and availability of on-the-ground resources means groups are able to carry out more effective interventions. Some partnerships are primarily driven by market factors, such as collaborating with organizations and individuals who specialize in or are familiar with issues in specific countries of interest to the platform. Others are motivated by particular issues, such as hate speech and antisemitism (topics of particular interest in the U.S. and Europe).¹²⁸

Apart from inquiries into what determines the geographical coverage of specific interventions by platforms and the development of emerging partnerships, other questions should focus on the uptake of technology in specific regions and the user experience. Tech companies and civil society groups aiming to intervene positively regarding PVE/CVE should be aware of who uses the technology and how they use it in specific locations. Would knowledge of the latter be informative enough for others who want to intervene in similar locations to replicate? Another important consideration pertains to data points confirming whether specific platforms are indeed being used for promoting terrorism in specific geographies (for example, TikTok might enjoy significant popularity among users in Africa, but it might remain unclear whether relative to other platforms it is being used by African users to propagate TVE). This kind of data could help with targeting interventions at the right platform to the right audience in the right geographic location.

Technology and Power

The study revealed that the majority of tech surfaces originate from the global North, with the remainder primarily consisting of Asian surfaces like TikTok.¹²⁹ What implications does this predominance of platforms from the global North hold for the future of positive intervention? Intervention efforts need to encompass considerations related to skills, knowledge sharing, and technology transfer in order to

.....
126 Alternative Platforms, BTWG.

127 Alternative Platforms, BTWG.

128 Alternative Platforms, BTWG.

129 Alternative Platforms, BTWG.

empower actors in the global South to establish and apply their own technological surfaces in a manner that aligns with their local context. This could be critical in emphasizing the salience of home-grown technology and how technology can be conceptualized to help advance certain cultural interests. This can be achieved by adopting specific designs that can influence users according to the intentions of those who own and produce the technology.¹³⁰ To avoid romanticizing its universal capabilities, it becomes important to investigate (for instance) how the application of technology across geographical landscapes would affect positive interventions in varying contexts. For example, how would positive interventions on TikTok in China be different from those in Africa, given that Facebook and Twitter are more popular among adults on the continent while WhatsApp and Instagram are favored by youth.¹³¹ These are questions that require comparative studies, which are unfortunately beyond the scope of the Blue Teaming Working Group output but merit exploration going forward.

Centering the Audience

One promising opportunity that arises from the study connected is how social media surfaces prioritize the experience of the audience/user. Even though this is an almost intrinsic consideration by virtue of the fact that social media surfaces thrive on User Generated Content, social media surfaces are increasingly embracing liberal/democratic/open technological designs by incentivizing the use and operation of their platforms. For example, they have help centers, conduct mentorship programs, advocate for best practices, and endeavor to devise ways of flagging harmful content/extremist content to reduce online polarization.¹³²

Centering the audience is a pivotal aspect of the broader philosophy of “tech for good” or “do no harm” mantras. As tech companies continue to dive deeper into innovating their surfaces to positively intervene in preventing violent extremism, it is critical that audiences are involved in research and development. Their participation helps researchers understand the phenomenon of local content moderation (which is gaining prominence) and rehumanize social media platforms across different geographical spaces, ethnicities, cultures, and contexts/locations.¹³³ A certain level of attention should focus on identifying who the users are and understanding their cultural orientation. If most of these surfaces are from the global North, what does this mean globally if (for example) gaming holds little significance in some global South contexts? How can positive interventions utilizing gaming platforms benefit users on a global scale? How can users/clients participate in platform research and development? One approach could involve incentivizing and training users to flag content online, given that platforms may have limited capacity for rigorous moderation.¹³⁴ This can also be facilitated through auto-detection and potentially

.....
130 James Lewis, “Technology and Power,” Center for Strategic and International Studies, March 30, 2020, <https://www.csis.org/analysis/technology-and-power>.

131 Ndlela and Mulwa, “Social Media, Youth and Everyday Life.”

132 Alternative Platforms, BTWG.

133 See Minna Ruckenstein and Linda Turunen, “Re-humanising the Platform: Content Moderators and the Logic of Care,” *New Media & Society* 22, no. 6 (2019): 1026–1042, <https://doi.org/10.1177/146144481987599>.

134 Ruckenstein and Turunen, “Re-humanising the Platform.”

by gamifying reporting by users.¹³⁵

Conclusion

As extremism mutates online, its complexity is compounded by how algorithms amplify negativity within the context of surveillance capitalism and the attention economy inherent in social media.¹³⁶ A more pragmatic approach calls for demystifying the complex nature of technology in everyday spaces found in the realms of “media in everyday life.”¹³⁷ This grants technology a central role in enabling human progress, especially in underprivileged societies. But for this to be realized, the conversation on how tech surfaces can be used for positive interventions in PVE/CVE needs to be expanded to encompass all segments of society. Often lacking in such conversations, policy makers and government representatives frequently exploit social media’s potential to spread their own versions of truth and state propaganda.¹³⁸ Civil society groups and practitioners continue to pioneer innovative methods to address online extremism using artificial intelligence-powered solutions such as the Peacemakers Corps Foundation Kenya’s (PCFK) Maskani.¹³⁹ Other digital peacebuilding movements (including micro-influencers who generate content) are significant players in arresting the commodification of hate and extremism and should also be part of the conversation.

Social media surfaces are a double-edged sword¹⁴⁰ that can be used for both conflict and peacebuilding. Careful attention should be given to their application to ensure they do not cause greater harm than good.

.....
135 Alternative Platforms, BTWG.

136 Lisa Schirch, “The Tectonic Shift: How Social Media Works,” in *Social Media Impacts on Conflict and Democracy: The Tectonic Shift*, ed. Lisa Schirch (New York: Routledge, 2021), 1–21; Fredrick Ogenga, “Social Media Literacy, Ethnicity and Peacebuilding,” in *Social Media Impacts on Conflict and Democracy: The Tectonic Shift*, ed. Lisa Schirch (New York: Routledge, 2021), 216–234; Eman El-shebiny, “Social Media Impacts on Civil Society, Violent Extremism and Government Control,” in *Social Media Impacts on Conflict and Democracy: The Tectonic Shift*, ed. Lisa Schirch (New York: Routledge, 2021), 81–88.

137 See David Morley, *Media, Modernity and Technology: The Geography of the New* (New York: Routledge, 2007).

138 See “Countermeasures for Mitigating Digital Misinformation: A Systemic Review,” International Panel on the Information Environment, 2023, <https://www.ipie.info/research/sr2023-1>; “Platform Responses to Misinformation: A Meta-Analysis of Data,” International Panel on the Information Environment, 2023, <https://www.ipie.info/research/sr2023-2>.

139 Maskani is a digital peacebuilding movement in Kenya that uses social media (WhatsApp and Facebook) for addressing polarization online (especially political extremism driven by ethnic politics). It is currently being piloted by over 70 youth influencers in six public universities in Western Kenya and the surrounding communities. See “MASKANI’ Digital Peacebuilding Closing Workshop Report,” Center for Media, Democracy, Peace and Security, October 23, 2020, <https://www.kpsrl.org/sites/default/files/2020-11/Maskani%20Final%20Gathering%20Report.pdf>.

140 Schirch, “The Tectonic Shift.”

A long-exposure photograph of a bridge at night, showing bright, diagonal light trails from vehicles crossing it. The bridge's structure, including railings and support beams, is visible in the foreground and background.

Copyright © Global Internet Forum to Counter Terrorism 2023

Recommended citation: William Allchorn, Samantha Kutner, Ellie Rogers, and Fredrick Ogenga, Playbook on Positive Intervention Strategies Online (Washington, DC: 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