Studies in Conflict & Terrorism

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uter20

New Models for Deploying Counterspeech: Measuring Behavioral Change and Sentiment Analysis

Erin Saltman, Farshad Kooti & Karly Vockery

To cite this article: Erin Saltman, Farshad Kooti & Karly Vockery (2021): New Models for Deploying Counterspeech: Measuring Behavioral Change and Sentiment Analysis, Studies in Conflict & Terrorism, DOI: 10.1080/1057610X.2021.1888404

To link to this article: https://doi.org/10.1080/1057610X.2021.1888404

Published online: 30 Mar 2021.

STUDIES IN CONFLICT & TERRORISM
https://doi.org/10.1080/1057610X.2021.1888404

Routledge
Taylor & Francis Group

New Models for Deploying Counterspeech: Measuring Behavioral Change and Sentiment Analysis

Erin Saltman[a], Farshad Kooti[b]*, and Karly Vockery[c]

[a]Director of Programming, Global Internet Forum to Counter Terrorism, London, England; [b]Senior Data Scientist, Facebook, Menlo Park, California, USA; [c]Online Safety Threat Analyst, Facebook, Menlo Park, California, USA

ABSTRACT

The counterterrorism and CVE community has long questioned the effectiveness of counterspeech in countering extremism online. While most evaluation of counterspeech rely on limited reach and engagement metrics, this paper explores two models to better measure behavioral change and sentiment analysis. Conducted via partnerships between Facebook and counter-extremism NGOs, the first model uses A/B testing to analyze the effects of counterspeech exposure on low-prevalence-high-risk audiences engaging with Islamist extremist terrorist content. The second model builds upon online safety intervention approaches and the Redirect Method through a search based "get-help" module, redirecting white-supremacy and Neo-Nazi related search-terms to disengagement NGOs.

ARTICLE HISTORY
Received 10 October 2020
Accepted 13 December 2020

Over the last 10 plus years there have been huge global efforts and investments in the Preventing and Countering Violent Extremism (PVE/CVE) sector. This is particularly the case since 2014 with global awareness of the digital savvy nature of the so-called Islamic State and their international recruitment strategy. Large waves of government and private funding have subsequently worked to support organizations, activist networks and marketing teams trying to counter online propaganda and recruitment. This form of strategic communication has largely been done through online messaging campaigns, commonly referred to as "alternative narratives," "counter-narratives" and/or "counterspeech."

However, the golden question remains as to whether or not these programs, local or global in their reach, have had a positive measurable impact on the audiences they aim to reach. Beyond measuring the basic metrics of reach and engagement, can these primarily online efforts show behavioral change and/or sentiment shift in the intended target audience exposed to this content? Could exposure to counterspeech in at-risk or radicalized audiences perhaps have the unintended consequence of further radicalization, or act as a catalyst to the radicalization process? How best can private tech

CONTACT Erin Saltman erin.saltman@gmail.com
*Present address: Dr. Farshad Kooti, 1 Hacker Way, Menlo Park, California 94025, USA. Email: press@fb.com.

companies work with non-governmental organizations (NGOs) and experts in the PVE/CVE space? This article aims to analyze these questions through two testing models of counterspeech deployment.

For the purposes of this article, we will refer to online PVE/CVE campaigns and initiatives as "counterspeech," with the understanding that this type of content can include both preventative or alternative messaging as well as more down-stream and directed counter-messaging. This article explores two approaches for measuring theories of change in audiences exposed to counterspeech that go beyond basic reach and engagement metrics, based on partnered research and testing carried out on Facebook. Testing and analysis were led by the Facebook Counterterrorism and Dangerous Organizations Policy Team in coordination with internal data scientists, community integrity engineers, market teams and safety researchers with privacy and legal reviews of both methodologies before their deployment.

Privacy and legal reviews have existed at Facebook for many years and are needed in order to launch any new on-platform program and tooling and have become even stronger in recent years. Privacy Review at Facebook includes technical validation, internal and external consultations and a review of documentation and implementation procedures. Review processes look closely at any usage of data, transparency measures for the user interfacing, assess risks and put safeguards in place to address any concerns before a program or tooling feature is allowed to launch. Internal stakeholders involved in Privacy Reviews include key reviews from product, policy, communication, design, legal, user experience and research team leads – among others depending on the nature of the project.

Both models were launched and tested in partnership with NGOs, who developed the counterspeech content and programmatic efforts independently of work with Facebook. The ultimate purpose of this research is to assess the efficacy of different counterspeech and counter-extremism efforts online in order to test the viability of methodologies for proactively surfacing counterspeech content to identified "at risk" audiences at scale and to provide constructive interventions. The first model uses an A/B testing format, building off of campaign deployment traditionally done through ads marketing tools. The A/B test focused on testing both softer preventative content and harder counter-content in order to assess behavioral change aimed at audiences with early at-risk indicators around violent Islamist extremist content in English and Arabic. The second model builds upon existing online safety intervention models that Facebook has developed and combines an approach with the Redirect Method to measure the effectiveness of online searches that trigger support toward offline disengagement practitioners.[1] This redirect initiative aimed at audiences searching for white supremacy and neo-Nazi groups and/or individuals and was initially launched in the USA, again with a localized NGO partnership.

The aim of presenting the results of these two methodologies is not to compare them with each other, but to explore the efficacy of two different approaches in disseminating counterspeech online to at-risk target audiences and see where, and what type, of impact is measurable. Therefore, while the violent extremist ideologies and geographies targeted in each model are different, that should not detract from the ultimate goal of testing new models for deploying counterspeech. Recognizing violent extremism takes different forms around the world, we chose to focus on countering violent Islamist

extremism and white supremacy/neo-Nazism trends as they are currently recognized by international researchers and international government bodies as the two largest violent extremist threats with an online nexus.[2] A range of other Facebook programs tackle a broad international scope of violent extremism.[3]

Based on the findings from testing both models, this research article argues that:

1. There is no data to support the theory that exposure to counterspeech negatively impacts behavior or sentiment indicators. There was no data showing "increased radicalization" based on either test.
2. Based on the A/B test, behavioral change in the wider target audience was not statistically significant overall, however, among the higher risk population identified within the target group, positive shifts of decreased engagement with violent extremist content were notable.
3. Focusing on behavioral signals to define an at-risk audience more accurately ensures that campaigns deployed are reaching the intended audience, thereby cutting out the noise around results and ensuring cleaner data for analysis.
4. One-off signals of a shared piece of violent extremist content are not clean indicators for an individual's sympathies or membership in a violent extremist group.
5. Based on the Redirect Initiative model, we have initial results of exponential increases in audiences finding online resources leading them to off-platform engagement with disengagement practitioners and resources. The findings show that an initial passive online search can lead to active offline engagement and initial findings of positive sentiment shifts.

Background Literature and Previous Research

Terrorist and violent extremist content online is not a new phenomenon,[4] however, its efficacy and ultimate role in the process of radicalization remains debated. Many researchers have argued that passive consumption of terrorist or violent extremist content plays less of an overall role in the process radicalization.[5] These studies have found that interpersonal connections and community influences remain crucial in effective radicalization. In contrast, others have argued that violent extremist content can, in fact, be successful in alluring and radicalizing in and of itself. This is particularly the case when the content targets specific audiences by attracting wider publicity, generating engagement and connecting with subgroups such as women and youth.[6]

While the debate on the exact role online content plays continues, it is generally agreed that the nature of terrorist and violent extremist content is diversifying; becoming increasingly interpersonal, targeted, cross-platform and global in nature. Therefore, the counterspeech content attempting to provide alternatives must also evolve to more accurately identify and engage personalized audiences, as discussed by experts and practitioners in this field.[7] In developing an approach to create and deliver counterspeech content online, it is important to remember that, if not done sensitively and appropriately, counterspeech could possibly cause harm or further alienation to an at-risk audience.[8]

With these factors in mind, counterspeech content should take into consideration not only the counter message itself but also the alignment of who the content is intended for, who delivers the message, what imagery or symbols are packaged around the message to give it authenticity, and what platform(s) the message is distributed on.[9]

Background Research in Developing the Two Methodologies

The concept development for the two testing methodologies was developed based off of a series of in-depth and semi-structured interviews conducted by the Counterterrorism and Dangerous Organizations Policy team in coordination with the Safety Research team. Interviews were held with a range of civil society and NGO counter-extremism practitioners as well as former extremists April through May 2017.[10] The aim of these interviews was to get expert and practitioner feedback on where there were existing gaps in social media approaches to PVE and CVE efforts. The interviews also explored tactics that would be appropriate and inappropriate for tech companies to deploy.

The ultimate issue that was brought up in almost every interview was that there was no existing internal proactive (mechanized) effort to challenge or redirect processes of radicalization on Facebook toward existing counterspeech content or resources. In other words, despite having internal indicators and strong program development via partnerships with NGOs who make counterspeech, the two remained largely separate. Facebook uses internal indicators to help surface, review and remove known violating content. Facebook also has a wide range of localized programmatic efforts with partners that support them in the development and deployment of counterspeech.[11] However, these tend to run in parallel, rather than in-sync.

The interviews also reiterated some long agreed upon premises and cautions around deployment of counterspeech:

- Counterspeech content should be localized in order to resonate with an intended audience. Facebook is not the credible voice of counter extremism efforts, but local networks, community practitioners and experts can be.[12]
- The content being surfaced should match the type of violent extremist ideology being countered.[13]
- The source or reason for triggering content should be transparent if it is being surfaced through Facebook mechanisms, and the mechanization for surfacing content should adhere to privacy concerns.[14]
- Counterspeech efforts that aimed at broad at-risk cohorts in a preventative capacity (like "the youth" in a given country or city) could be reached relatively effectively through wider ads marketing tools that optimize targeting of 100,000 plus audiences. However, it was extremely difficult, if not impossible, to reach low-prevalence-high-risk audiences in a targeted counter-extremism approach to ensure content was reaching those on an early spectrum of radicalization.

The Facebook Counterterrorism and Dangerous Organizations Policy team began developing the two methodologies discussed in this article in order to address the gaps based on the feedback from expert practitioners. Looking at the below lifecycle of

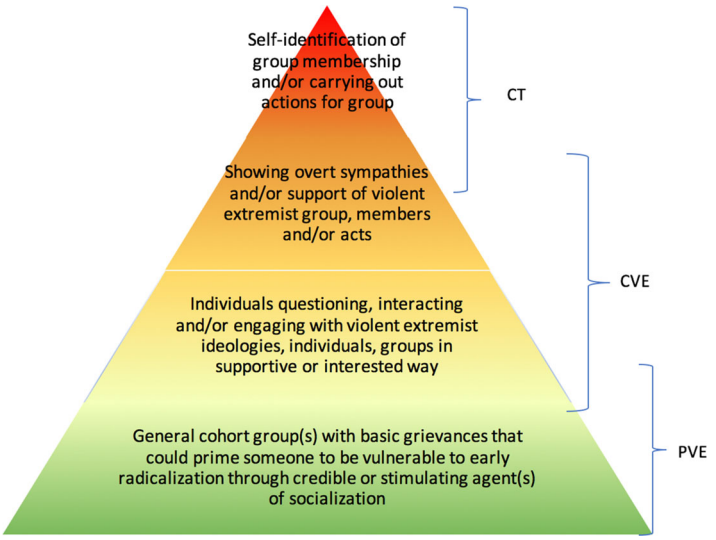

Figure 1. To the right of the pyramid are the potential reactionary frameworks for engagement; Preventing Violent Extremism (PVE), Countering Violent Extremism (CVE) and Counterterrorism (CT). There are areas for overlap in each response level depending on the available online "symptoms" or indicators available to assess radicalization. The two methodologies address the CVE space using hard indicators to trigger soft responses in combination with the removal of any violating content.

radicalization, the gaps remain within the targeted CVE space (Figure 1). As an individual undergoes a process of radicalization toward violent extremism they go from a banal, non-threatening individual that could be considered "at risk" in a prevention framework, into being considered a risk to others and at risk of more violently inclined indoctrination. The pyramid highlights the general process of radicalization, while recognizing that this process is not necessarily a linear one, and that an individual might turn away from or become disinterested in a given violent extremist group or ideology for a myriad of reasons. The pyramid is based on interviews with counter-extremism and counterterrorism practitioners while building on a model put forward by ISD Global.[15] However, it is important to note that as the process of radicalization continues, nearing more toward violence, fewer individuals will be a target cohort. Alternatively, an individual can remain on any given level, or even turn away from the radicalizing process due to a number of personal and external socializing factors.

Methodology 1: A/B Testing Proactive Surfacing of Counterspeech in English (U.K.) and Arabic (Iraq)

The goal of this methodology was to develop a model whereby a hard indicator of engagement with a violent extremist group or piece of content could trigger an automatized soft response, by surfacing a set amount of relevant counterspeech to that audience, over a specified period of time. The main question being asked is whether exposure to counterspeech content has an effect on a target audience's continued engagement and/or consumption of violent extremist content. In essence,

whether or not a noticeable behavioral or sentimental shift can be measured. This was done through two quantitative analysis models and two qualitative analysis approaches.

Quantitative analysis focused on the aggregated changes in the behavior of the control group in comparison to the treatment group. We compared the rate of violations after the first exposure to the counterspeech videos and for the following 90 days in order to see if we could detect changes in behavior when it came to how often someone may violate the platform's content policies. From our earlier analysis of time between violations, we know that the 90 days' time period includes more than 99.9% of future violations that are likely to occur. Besides the propensity and likelihood for further violations, we also analyzed the statistics on engagement with the counterspeech content and viewing activity.

Qualitative analysis aimed to understand the before and after behavior of the target audience population through a few deep-dive analyses around a range of activities. Activity analysis included seeing if there were on-platform for violation trends, if a user had trends in abusing Facebook community standards in different ways. It also included assessing the discourse in comments below posts, how many people viewed content that violated Community Standards before it was removed, and potential connections with other users that may have been removed from the platform violating policies against violent extremism. In essence, this qualitative analysis was designed to build on the quantitative results in order to analyze more nuance in the behavioral trends of the target audience.

There are also some key questions that need to be answered when testing the impact of counterspeech exposure before any testing methodology can begin, which interviews with experts highlighted: Who is the target audience? What is the message being conveyed to the audience? Is the messenger credible? What is the rate and frequency of counterspeech content exposure needed to see a shift in sentiment? What is the measurement and evaluation framework needed to show sentiment and/or behavioral shift as a consequence of counterspeech exposure? With these in mind we developed our A/B testing model.

A/B testing is a common methodology for statistical hypothesis testing, where two variants A and B are tested, and the results are compared.[16] In this case, our hypothesis is that exposure to counterspeech will reduce likelihood of further production and/or engagement with further violating terrorism content. To test this hypothesis, first we randomly assign people to two groups with the same size and show the counterspeech to one group, i.e. "treatment group," while we do not change anything for the other group, i.e. "control group." Then, we compute the average rate of violations for each group. As there could be some differences just due to chance and noise, we have to make sure the differences are *statistically significant*. To test the statistical significance, we compute 95% confidence intervals of the mean of violations rates.[17] The 95% confidence interval is a range of values that we can be 95% certain that it includes the true mean of the population. If the confidence intervals of two groups do no overlap there is a statistically significant difference between the treatment and control groups, otherwise the A/B test is not showing a statistically significant difference between the two groups.

Target Audience

Generally speaking, most counterspeech still struggles to show that it is truly reaching its "target audience." Most counterspeech globally has been launched online by one of three ways: by organically distributing content via existing grassroots networks, using targeted ads campaigns, or by triggering counterspeech "playlists" of content in the results yielded by specific search terms (the Redirect Method). There are benefits and limitations to all three approaches. Tapping into existing or known networks online means that there is a better chance of peer-to-peer credibility for the content to land, but it is harder to reach those vulnerable individuals outside of your network. Targeted ads campaigns can be used to reach audiences based on their likes, interests and behaviors, which can potentially penetrate into harder to reach online insulated communities but there is no control over the rate and frequency of exposure, meaning content might land as a one-off within someone's social feed. Lastly, triggered playlists can present a range of counterspeech content to a user but there is no way of knowing why an individual has searched for a particular term (are they actually a sympathizer or perhaps a journalist, researcher or activist?).

For the purposes of this A/B testing methodology, we wanted to surface counterspeech content to individuals that showed a hard indicator of interest or engagement toward a violent extremist organization. It was also important that the counterspeech content surfaced was specific or related to the violent extremist ideology. This audience type is what we would call a "low prevalence, high risk" group. Meaning, there is not a huge audience of this nature, relative to Facebook audience size, but they are showing hard indicators of being at-risk of, or vulnerable to further radicalization. Our audience for this testing consisted of individuals who had early indicators of consuming, sharing and/or engaging with larger global Islamist extremist terrorist movements, like the so-called Islamic State or Al-Qaeda propaganda content.[18] This audience had not violated Community Standards to the extent they should be disabled, but engaged with violating content posted by others that was then removed.

In order to have a more robust result, and to avoid bias that might be found in one language or geography, we A/B tested in two language markets. We focused on English (U.K.) and Arabic (Iraq) due to country size for measurement, and each country's recent history with Islamist extremist terrorist incidents. Half of those identified as applicable for "target audience" typology were surfaced content, while the other half were left as a control group to observe normative behavior without counterspeech exposure during the same period. During the two-week duration when the target audience was identified and surfaced content to, a total of 37,000 individuals were included in the test group, receiving in-feed counterspeech content in the two country markets. The same number of individuals were identified and not exposed to counterspeech content for the purposes of A/B testing.

Messages and Messengers

Just like in broader marketing optimization, in counterspeech, both the message and the messenger have to be credible for the target audience for counterspeech to resonate with an audience. As discussed in the introduction, studies on PVE/CVE work have

shown that credible messages need to come from some element of credible messenger in order to have cognitive impact. With respect to this study and the audiences engaged with for this purpose, Facebook was not going to be the credible messenger. Just like documented difficulties governments have had with credibility in messaging within the counter-extremism space, private companies are coming from a similarly difficult position of having little credibility being the originating voice of the counterspeech. Therefore, for the content surfaced to resonate, it has to come from third parties that are credible either due to the nature of the content itself and/or due to who the NGO or messengers within the context are.

Part of the reasoning for choosing the two country/language markets was because we had trusted NGO partners known to us, making counterspeech content suited for both. We wanted to ensure that we had two content partners for each market; one focusing on more upstream lighter preventative content and one focusing more downstream on hard hitting content to compare engagement and metrics across the spectrum. One of our partners produces content in English and Arabic fluently for multiple country-markets, making them a viable partner for both the U.K. and Iraq. This was paired with a secondary NGO partner for each country-market. All three NGOs already had content they put out on Facebook. They also already had used Facebook ad credits and tooling to amplify their messaging online, so they were relatively familiar with Facebook-specific tools and processes.

The International Center on Security and Violent Extremism (ICSVE) was the first NGO we partnered with for this A/B Testing. ICSVE is a U.S.-based, but globally engaged NGO focused on counterspeech messaging, primarily challenging Islamist extremist and jihadist recruitment narratives. ICSVE produces content they launch in a range of different languages on Facebook, YouTube and Twitter. Having run basic metrics through previous ads campaigns with ICSVE, we knew their videos received high reach and engagement through ads and wanted to sample their more hard-hitting content through this A/B test. The video content of ICSVE had already been successful in a range of language markets. ICSVE has content fluently in Arabic and English, so served as a test case for both markets and represented the more "downstream" testing content. Downstream refers to overtly hard-hitting and openly challenging Islamic State. The content used from ICSVE were edited videos of former Islamic State fighters telling their stories of both why they joined and why they left or regret their decision. Dr. Anne Speckhard, who leads ICSVE, is a trained psychologist who has conducted hundreds of interviews with former Islamic State members within prisons or in rehabilitation centers.

For each language, we wanted to pair the harder-hitting content of ICSVE by testing more upstream, preventative or resiliency building content. For the English language market, we partnered with ConnectFutures. ConnectFutures is a U.K.-based NGO that has many years of experience in the prevention and resiliency space. They work with local schools and institutions across the U.K. to challenge far-right and Islamist extremist ideologies through positive and alternative narratives. They provide advice and have developed toolkits for a range of sectors on proactive and sensitive ways to engage primarily youth audiences around the topic of extremism.[19] They are also a member of the Online Civil Courage Initiative (OCCI) - the Facebook initiative run with the Institute for Strategic Dialogue.

For providing the preventative, resiliency content for the Arabic language market in the A/B test we partnered with The Adyan Foundation, based in Lebanon. On its "Taadudiya" online platform, Adyan has created a range of Arabic language and Iraqi specific content over the years. They work closely with partners across the Middle East, and are internationally recognized for their alternative online messaging, which they call "existential narratives," providing content that promotes resilience, compassion and diversity. Adyan works on valuing cultural and religious diversity in its conceptual and practical dimensions, and on promoting coexistence and diversity management among individuals and communities, on the social, political, educational and spiritual levels.

With all three NGOs, it was important that the content itself was independent. None of the content was funded by or developed specifically for Facebook. The content was developed for the purpose of engaging with audiences that were sympathizing with Islamist extremist narratives locally with the intention of planting a seed of doubt, providing an alternative narrative or directly undermining the recruiting tactics of the so-called Islamic State. The A/B test also juxtaposed "hard" and "soft" content. While ICSVE provided more obvious content explicitly countering Daesh content, Adyan Institute and ConnectFutures provided softer localized content about resiliency, interfaith dialogues and community engagement. Each NGO provided between seven and ten pieces of content to test with.

Rate and Frequency

One of the largest unanswered questions about the efficacy of counterspeech in counter-extremism approaches is about how much counterspeech you would need to expose an audience to, and how frequently, to see a change in sentiment. There has been previous research showing that well-constructed counterspeech has the potential to go further online than ill-constructed hate speech or extremism messaging.[20] There has also been a range of research that talks through what makes counterspeech, or counter-extremism messaging, most effective.[21] However, there is no framework to assess rate and frequency of counterspeech content on a viewer in order to achieve a sentiment impact. To answer this specific methodological question, in order to build a pipelined framework for counterspeech surfacing, Facebook worked with a group of master's students at Stanford University.[22]

In this study, the Stanford Masters students developed an innovative and flexible research design, tested with in-depth qualitative analysis (pre- and post-exposure surveys) using a 100-person test audience. In this design, they gave an attitudinal survey using openness or aversion to gay marriage, with Australia as a case study since the country was questioning gay marriage rights at the time. They then had participants come in to view a range of both pro (counterspeech) and anti (hate-based) gay marriage content which came from real world examples online and in news media, but which had been anonymized for the study. They changed the rate and frequency of counterspeech exposed to different segmented audiences among participants. Participants were engaged for 4, 7 or 10 days.

Fluctuating days of exposure and quantity of counterspeech shown per day, the research group were able to analyze variation in sentimental shifts based on rate and

frequency. The results of this research showed that at least two to three pieces of content a day were needed to ensure the message was conveyed and reinforced. Messaging needed at least 4 days for any notable sentiment shift to be apparent. Messaging for 7 days or more did not have any significant impact on sentiment shift than that seen after at least 4 days. Based on this guidance, it was decided that the A/B testing should surface 2 to 3 pieces of content per day to target audiences for a period of 5 to 7 days.

Mechanism for Surfacing Content

To date, the vast majority of counterspeech has been launched one of three ways; (1) through organic existing networks, (2) through ads marketing tooling or (3) through search term redirection. All three methods have progressed putting forward a myriad of messaging to counter hate and violent extremism online, but the measurement and evaluation of these approaches is largely based on reach and engagement, where exacting rate and frequency analysis is not possible and sentiment or behavioral change analysis is near impossible. Organic networks can provide greater legitimacy of source, but usually are limited to existing sympathizers to a cause. Ads can reach an approximated low-prevalence-higher-risk audience type if done strategically but cannot ensure any specific list of individuals has the content surfaced, often getting caught competing against other competing ads. Search redirection, whereby a given term or phrase list in a platform's search field yields counterspeech content, instead of extremist content that the search was intended for, results in a closer approximation in targeting an audience, and can measure whether an individual has seen more than one of the intended counterspeech within a playlist but is based on one-off search interactions.

Due to the above needs, usage of Facebook Ads are less sufficient for accurate targeting of smaller specialist audience types. Ads tooling is built for widest audience optimization possible and does not optimize for low-prevalence/high-risk audience types. Ads tend to be best optimized when you target 100 K + audiences, while this testing was aimed at under 50 K. Ads are also unable to ensure that a specific and refined target audience is definitively served the content, as it competes with other ads variables, and there is no way to ensure a small series of content in a given period of time. As with the identified issue with many of the existing evaluations around counterspeech to date, ads do not allow for evaluation beyond surface level metrics of reach and engagement.

Thus, for the purposes of this A/B test model, we decided to use Quick Promotions as the most efficient way to ensure that a certain rate and frequency of content was surfaced to a specified audience. Quick Promotion is a framework that allows Facebook to communicate directly to users with well-targeted, well-timed information. Quick Promotions also allow for; guaranteed delivery to an audience, ability to better track users click-through-rates, video and photo could be formatted within the frame, clickable link to third party and/or internal content, translatable to different languages and a customizable area for text.

It is important to note that Quick Promotions have their own guidelines to ensure full transparency to the audience. Transparency that the promotion was triggered by Facebook was given in the frame of the content, and transparency about the NGO delivering the message was also given. The partnership between Facebook and the NGO partner was also made clear. Understandably, this could both help and hinder potential

Figure 2. Above shows the mock example of how content was surfaced to individuals through Quick Promotion formatting within the A/B testing treatment. It is visually distinct from organic content in a Facebook newsfeed, for transparency purposes. Facebook gives an indication of why the content is surfaced and is open about a partnership with a specific non-governmental organization, whose content is surfaced within the Quick Promotion.

effectiveness of the counterspeech campaign on a given audience. On the one hand, because it is formatted differently than normal organic in-feed content, and differs from an ad, it could benefit the model by attracting the audience to look at it longer due to differentiation. On the other hand, given research into at-risk audiences and potential violent extremist sympathizers, it could also dissuade the targeted audience from engaging with the content, because of wider skepticism of Facebook surfacing the content and its partnership with a given NGO. Either way, the decision was made to ensure any content optimization by Facebook was given transparency, as guided by privacy and legal reviews around the methodology, as discussed in the introduction.

Importantly, for both models discussed in this article, no Personally Identifiable Information (PII) was shared to third parties. The NGOs partnering with Facebook who created counterspeech content were not given preferential access to internal information, accounts or data beyond what the average Facebook user can already see. Facebook imposes strict restrictions on how its partners can use and disclose the data provided. Figure 2 shows an example of how content was surfaced to individuals through Quick Promotion formatting within the A/B testing treatment.[23]

Results and Findings

Analyzing the results of the A/B test yielded neutral audience behavioral change upon broader quantitative analysis, but positive audience behavioral change when the

audience was further parsed out to focus primarily on higher risk audience types. In total, 74 K individuals were identified through our initial audience criteria, and half underwent the counterspeech module so that we could A/B test between those who had been exposed to counterspeech and those that had not. Thus, 37 K individuals went through the counterspeech model. During the four weeks that the module ran, the ratio of U.K. to Iraq audience was roughly one to five. This divide has multiple potential reasons having to do with popularity around certain types of content and/or groups as well as current events during the time the module ran (23 October–20 November 2019). The average age of individuals fitting the criteria for A/B testing was 29-years old as indicated on a user's profile. There was a significantly larger proportion of men fitting the at-risk audience criteria compared with females. Male audiences made up 87.4% of the sample.

It is first necessary to note how much of the audience watched the surfaced content in any measure that could be recorded, rather than scrolling by it without pause. A high rate of the audience was ensured viewing at the most basic levels. 81% of the test group watched at least one of the videos within the counterspeech series that surfaced, and there was a large amount of variation in how much a given individual engaged with the series of content and how much they watched of any given video. While the Quick Promotion format ensures that a video appears in an audience's newsfeed, with the relevant transparency framing, it cannot force a user to watch the content more than any other piece of content. Such is the same with ads, or any organic content shared by friends or social networks an individual is connected with on Facebook. As a result, the viewership of the counterspeech content was fairly low. From the total possible 40+ min of view time, only 23% of the users watched more than 1 min in aggregate, and only 5% watched more than 3 min. While this might make the effects of the treatment weaker, in fact this is an important finding of this study that voluntary viewership of the counterspeech content does not result in high levels of engagement.

Like most other counterspeech models, this highlights an understandable limitation to the A/B test, even with this more precise mechanism for surfacing counterspeech content at a given rate and frequency. Unlike the Stanford University study, where a sample audience gives permission to have counterspeech surfaced to them, and ensures full viewership of each piece of content, in this model we can only ensure the surfacing of the content. The audience is not being forced to watch or engage any content in full or in part. Therefore, and as noted by other research around the effectiveness of counterspeech, the content itself has to be compelling in and of itself to the audience regardless of the specialized targeting method.

The comparison of violation rates between the control and treatment groups did not yield statistically significant grounds to claim that the test model positively changed the wider target audience's behavior as a whole. There were no significant total changes to the rate of re-abuse on the platform, nor further changes to trends of views/engagement with violent extremist content that was later removed from the platform. However, that is not to say the quantitative analysis did not highlight important key lessons. In this context, two important takeaways were noted, helping guide our qualitative deep dive.

First, the quantitative analyses did not show significant indication that the exposure to counterspeech led to increased engagement with violent extremist content, nor to

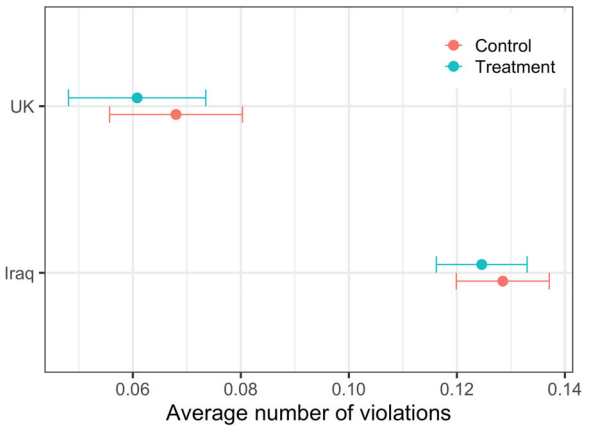

Figure 3. The above shows a comparison of violation rates for the users in the U.K. and Iraq, between the control and the treatment groups. Average number of violations along with 95% confidence intervals do not show any statistically significant differences between the control and treatment groups.

any catalyst in indicators for processes of radicalization, debunking some concerns about wide distribution of counterspeech. Figure 3 shows the average number of violations within 90 days after the exposure for the control and treatment groups, for the U.K. and Iraq, along with 95% mean confidence intervals. Our results do not show a statistically significant change in the rate of the viewership of content taken down for terrorism reasons, because the confidence intervals for control and treatment groups overlap. That being said, despite a much more concise model for identifying at-risk audiences through hard indicators - such as views, shares and or engagement with violent extremist content - there was a large amount of variation in the audience typology.

However, when the segmented audience was reviewed more closely, it was apparent that there were three broad sub-categories within the A/B test audience. The qualitative research team took a deeper analysis into 23 cases profiled as low risk (3 profiles), medium risk (10 profiles) and high risk (10 profiles) audiences types based on how much or little the profile had viewed, engaged or previously shared violent extremist content that had been removed for violating Facebook's community standards before the A/B test started. We divided these profiles on a risk spectrum in the event that behavioral changes could be observed based on the quantity of previously recorded violations. Of these profiles, 22% were female and 78% were male, similar to the wider audience breakdown. Through qualitative analysis, the following three sub-categories become apparent.

False Positives: Broadly speaking this group consisted of individuals who might have shared one plus pieces of violent extremist content that had subsequently been removed, as per Facebook community standards. However, on closer review consisted largely of activists or more politically engaged individuals who were either openly condemning terrorist organizations or commenting on regional geo-politics and news. Thus, the hard indicator of highlighting an audience purely on sharing pieces of even known terrorist or violent extremist content, is in fact a very noisy audience with a range of individuals who would not likely be considered an at-risk audience on closer review.[24]

There was no evidence of any negative impact from surfacing NGO counterspeech content to this audience, however, they should likely be discounted from the overall review of whether or not the model was successful. This is also salient as some have suggested that shares of terrorist content should be linked to more proactive disclosures by tech companies to governments about those individuals sharing this content. Based on the results of this study, this would be a dangerous overstep of privacy to yield personal information of an audience that will consist of significant false-positives. The sole act of sharing terrorist content on a public platform does not provide credible indicators that the person poses a threat to public safety such that tech companies should proactively disclose information about the person to law enforcement at the expense of the person's privacy.

Mixed Noise: This sub-category might be sharing pieces of violent extremist content, but there is not obvious sympathy toward a singular violent extremist ideology. Within this grouping, behavior indicates that content sharing trends are mixed with wider engagement of graphic violence content, gore and/or conspiracy theories. It would be difficult to say that this group is definitively sympathizing with a terrorist organization or singular group but has broader fascinations with violence. This type of mixed noise group was suggested by many articles in the aftermath of violent extremist atrocities, such as the Christchurch attack in New Zealand (March 2019), whereby video copies of the attack, which were originally live streamed, were shared widely in clip or full format across news media but also heavily among broader audiences online.

Violent Extremist Engagers: This sub-category consisted of individuals with higher hard signals of views, shares and/or direct engagement with violent extremist content. This group was of significant interest, seeing as it more accurately depicted what the target audience was meant to be; those showing signs of sympathy toward or interest in a violent extremist ideology. Notably, qualitative analysis did indicate that among profiles of this nature, there were instances that after exposure to the A/B test series of counterspeech content, no further on-platform violations or violent extremist sympathies could be seen. In other words, there is some positive indication that among the most at-risk target audience, the behavior of the individual did in fact change to neutralize and move away from visible sympathies toward explicit extremist sympathies.

The explanation for this could be twofold. On the one hand, we could speculate that the exposure to a series of directed and curated counterspeech content appearing at a tested rate and frequency in the newsfeed did in fact produce an indication of positive behavioral change, moving away from the more extremist and/or violent sides of an ideology. It is also worth speculating that the nature of having Facebook transparently surface counterspeech into the newsfeed could trigger in the audience an awareness that they are being targeted for a particular extremist belief and thus could change how they post publicly to avoid being "on the radar." The Stanford research that rate and frequency were based on seems to imply the former hypothesis explaining behavioral change. Either way, decreasing the potential for further sharing and engaging of violent extremist content on platform has a secondary positive effect by limiting further spread of the violating content in future habits and thus preventing further exposure of the content to vulnerable audiences.

This first testing methodology speaks to a model attempting to more accurately measure behavioral change of a given audience, based purely on exposure to a series of

counterspeech content. Initial results show a net neutral to positive result in using techniques like this to reach low-prevalence/high-risk audience types. The second methodology looks to measure the potential for sentiment, specifically by turning a passive online engagement into an off-platform active engagement between audience and practitioner.

Methodology 2: Search Redirection "Get Help" Module

The first model discussed in this article tested on target audiences at risk of violent Islamist extremist indicators with a methodology aimed at testing behavioral change. In contrast, this model focusses on target audiences at risk of white supremacy and neo-Nazi indicators - which explicitly violate Facebook's community standards,[25] with a methodology aimed at testing sentiment change. The aim of this second methodology is to test the capacity of online partnerships between social media companies and disengagement NGOs in their ability to turn passive online content searches with extremist indicators on a platform, into potential active engagement between an at-risk audience and a trained outreach professional off of the platform.

As discussed in the introduction, it is still debated to what extent online content is a source or key component of radicalization, but it is largely agreed upon that the Internet can at least be seen as a catalyst for the process of radicalization and possibly causal, depending on how active, as opposed to passive, online engagements between a recruiter and recruit become.[26,27] Assuming that passive viewing and engagement with online content can be a gateway to more active engagements with a recruit or extremist group, this second methodology looks to replicate the model of passive searches and content leading to the opportunity for more active engagement with a localized online community that can facilitate disengagement from an extremist group or ideology.

Going back to original interviews conducted with counter-extremism practitioners in 2018, advice guided us toward methods where risk-mitigation approaches continued to use Facebook as the primary trigger and pipeline holding the key to on-platform hard indicators of interest in a form of violent or hate-based extremism. However, that indicator should trigger the surfacing of credible connections to localized, real-world help services. While the A/B testing model previously discussed aimed to surface a broader array of content with rate and frequency metrics, this model aimed to turn a potential passive search into active engagement.

This second methodology was set up with a more simplified, unidirectional model, combining learnings from tried and tested Facebook safety responses connecting search terms to localized resources, as well as learnings from the Redirect Method. In developing this model, we first took from learnings within Facebook that were seen in other safety spaces used to mitigate other types of at-risk audiences. Facebook already has a range of tools and user experience to mitigate harm and provide resources to audiences. This includes models to risk mitigate indicators around harm types such as, suicide and self-injury, drug abuse, bullying and harassment and child safety.[28]

As an example of how this model has previously worked, we looked at how Facebook set up "Get Help" modules around the opioid epidemic in the United States. In this case, Facebook committed to making it easier for people who struggle with opioid

misuse to find treatment and resources for recovery. Among other measures and partnerships, searches on Facebook and Instagram about buying or selling opioids, slang terms for finding drugs and coded language for procuring drugs (which all goes against Facebook's community standards) would trigger and direct people to the SAMHSA National Helpline Information Page and other resources that provide free and confidential treatment referrals.[29] This would show up at the top of search results around the related drug misuse terms list and Facebook could track the Click Through Rate (CTR) around various terms.

We combined basic learnings from the "Get Help" module approach with the "Redirect Method." The Redirect Method was developed by a range of NGOs in collaboration with Jigsaw. In its pilot research, the Redirect Method is a 4-step approach that employs readily available online counterspeech content and existing online videos with targeted advertising tools to counter ISIS recruitment efforts online. In essence, when you search on YouTube or Google Search for a word or phrase that implies an interest or sympathy toward ISIS, your search results will yield a playlist of counter content, or counterspeech.[30]

The Redirect Method was developed with search engine platforms in mind. Google Search, YouTube and Bing have all deployed similar tactics, recognizing that individuals use the search functions to find specific content or information. Generally speaking, Facebook's search functionality is utilized in very different ways. Facebook as a platform has a different mission statement and usage compared to search engines. Facebook's aim is to give people the power to build community and bring the world closer together. As such, searches on Facebook for on-platform content are inclined more toward searching for individuals and/or groups to connect and engage with, as opposed to wider Search Engine queries which connect individuals to content from across the web. Thus, we looked for a modified version of a redirection model, more in-line with previous Facebook "Get Help" modules, recognizing the type of engagement an individual was looking for through the search.

Partnership and Audience Interaction

In March 2019, this Facebook redirect initiative was launched in the United States so that when people search on Facebook for specific terms related to white supremacy or neo-Nazism, results trigger a Get Help module at the top of search results linked to Life After Hate, an organization that provides crisis intervention, education, support groups and outreach. Therefore, the hard indicator of interest and/or potential sympathies in a violent extremist group are based purely on searching for related entities. In the development and framing of this type of redirect initiative, it is crucial that the NGO partner is looped into the development of terms and framing of goals from the beginning.

Life After Hate is a very specific NGO because of how it was founded, its innovative program ExitUSA, and its internal expertise on processes of radicalization. Founded by former extremists, its mission statement is a commitment "to helping people leave the violent far-right to connect with humanity and lead compassionate lives."[31] Their primary goal is to interrupt violence committed in the name of ideological or religious beliefs. They do this through education, interventions, academic research and outreach.

Life After Hate operates its program ExitUSA, bringing innovation, best practices and research informed services to the disengagement support infrastructure to help people proactively leave white supremacy and neo-Nazi movements. They also provide support to family and social networks. As stated on their website:

"Founded and run by former hate group members who have led successful post-movement lives, ExitUSA provides support to individuals who are looking to leave racism and violence behind." The organization uses a variety of strategies, including public awareness campaigns, peer-led support groups and access to relevant resources and referrals, paired with staff who have specialized training related to working with far-right extremist populations. This includes leveraging strategic community partnerships to help individuals get their lives back on track and on their way to making positive contributions to society.[32]

In a Redirect Initiative of this nature, with one primary NGO partner, there is a fundamental difference to other counterspeech or redirect methods. Instead of introducing an individual, through whatever trigger mechanism, to a range of content, this approach introduces an individual to an organization that can provide tangible support to leave a hate-based movement. Counterspeech content is often passive, meaning it is there to be consumed and often does not have a clear follow-up action the consumer needs to take. That means that this methodology will always be limited to either metrics of reach and engagement, or that in the best-case scenario basic indicators of behavioral change can be identified (like in the A/B testing model presented in this article). By introducing an NGO with tangible help outlinks, with real world geographical attachment, the audience is presented with an engagement option with active participation.

We can therefore measure the willingness of an audience to engage with an NGO's help resources and measure behavior and sentiment in two ways. Behavior can be measured in the basic form of CTR from the Facebook search leading to an individual clicking on the Redirect Initiative Module. CTR indicates how many link clicks content has received on a given ad compared to how many impressions the ad received. It is a common metric used by online advertisers to understand how ads drive traffic to websites and other destinations.[33]

That CTR shows us which terms are most likely to lead to an individual's willingness to engage with a third-party NGO that is not Facebook and that is openly offering disengagement and counter-extremism support. From there, Facebook metrics stop. It is up to the NGO where that module lands and it is often on their third-party website where they have the clearest resources and ways to engage with practitioners. Once an individual clicks to the NGO, Facebook is reliant on the partnership with the NGO to measure increases in website engagement, increases in enquiries and outreach requests.

For that reason, it is important that the NGO is a trusted partner, that they have credibility or a longstanding history of working in the counter-extremism and disengagement space, and that they are willing to communicate results. Increases in individuals reaching out for help and getting the help they need to disengage with a violent extremist group or ideology is the hardest target audience to reach. This is where we can see a potential online passive search turning into structured engagement leading to real sentiment change.

Target Audience and Redirection Criteria

One of the biggest challenges in developing programs for safety related issues is translating the way Facebook interprets an indicator and transforming that into meaningful and effective user-facing content. In order to uphold Facebook's voice of being "simple, straightforward and human," it needs to be clear why a given term used in search might surface a help module. There is a great need to control the vocabulary and phrases used in the redirect initiative and be explicit about what specific terms mean.

The target audience is based entirely on their own action of searching for a given term that relates to a predeveloped list of terms and phrases intrinsically linked to white supremacy and/or neo-Nazi individuals, groups or coded phrases. For this methodology, it was important that the list was co-developed with Facebook and the NGO so that there was transparency over what terms would lead to the NGO being highlighted as the top search result, with transparency about why Facebook was highlighting this module. Facebook representatives from the Market Specialists, Safety and Dangerous Organizations Policy Team came together to input and review around the list. For this reason, it was important that for Phase 1 of testing, the list erred on the side of being concise and restricted. Terms and phrases had to link directly to the form of hate-based extremism the NGO is best equipped to provide resources around.

Therefore, while many terms might be disallowed for hate-speech or slurs based on Facebook's community standards, we discluded broader racist and xenophobic terms. This also recognized the nature of how individuals search for things on Facebook. The terms list focused on banned individuals and groups because Facebook is more often used to find individuals and groups. This list also incorporated some of the known slogans and slurs used more commonly as coded language among certain white supremacy and neo-Nazi sympathizers. It has been well documented among experts on far-right extremist movements that adversarial shifts within this group are often seen by incorporating coded language and symbolism.[34]

Lastly, after a terms list was co-developed by Facebook and the NGO, the Facebook operations, markets and policy team ran testing on the list against potential false positives, to concentrate the list down to its most concise form. In Phase 1, a few hundred terms were agreed upon to pilot the project across the United States.

Search Experience

The search experience for an individual in the United States searching for one of the terms on the predesignated list is meant to be straightforward. In essence, it is programed so that the search results yield an information box anchored at the top of the search engine results page with the Help Module leading to the NGO. Whether or not to click on the module is completely up to the individual that searched for the term. The module links to a new experience with content that has been curated by an NGO, providing resources at a pivotal moment to assist the searcher away from going down their intended pathway. Figure 4 is an example of how the module might appear to the individual.

Projects like this, as in the A/B testing model, have to go through a rigorous privacy and ethics review process, as discussed previously. It is important to reiterate that at no

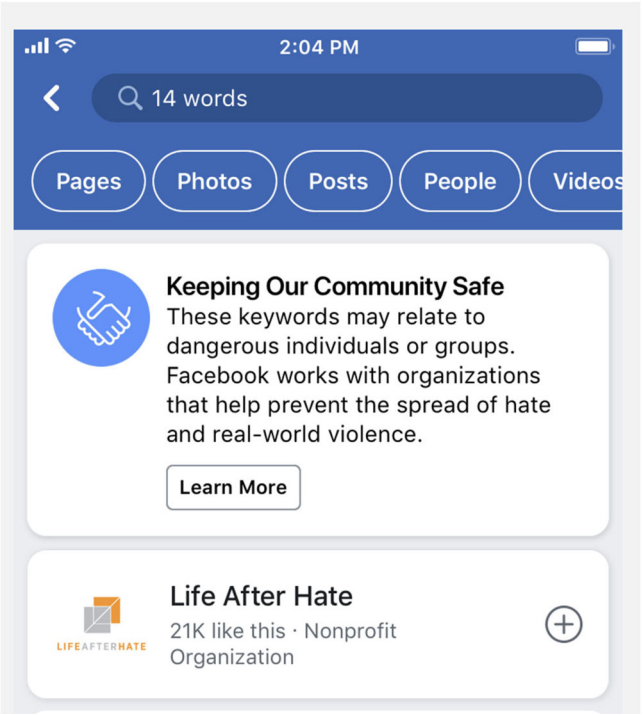

Figure 4. The above figure shows a mockup of how the module surfaces for individuals who have searched for specific terms that are part of the Search Redirect program. These are country and partnership specific.

time during this testing model was PII shared to a third party. The language that appears first is meant to be transparent from Facebook to indicate why this help module is being displayed:

These keywords may relate to dangerous individuals or groups. Facebook works with organizations to help prevent the spread of hate and real-world violence.

This is followed by simply introducing the NGO as a nonprofit organization and letting the individual decide on whether or not to click onto the module. Once clicked on, the NGO decides the journey pathway. In this case, and in others, Life After Hate decided it was best to have the module land onto their page that had a range of counterspeech videos, primarily featuring testimonials of former extremists that had left white supremacy and neo-Nazi movements. This was paired with a more featured button at the top of this page for getting help so that it was clear that the individual could trigger an interface with the NGO. At this point, the process, metrics and diagnostics - such as Google Analytics - is blind to Facebook. The reliance is on the NGO to communicate what they are seeing as results on their side.

Results and Findings

There are two aspects to collecting results to measure and evaluate the effectiveness of this project so far. The first are by looking at Facebook's metrics around CTR and

Engagement (lifeafterhate.org/media)	2018	2019												2020	
	Dec	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sept	Oct	Nov	Dec	Jan	Feb
Unique Pageviews	223	288	246	751	265	383	496	759	1,330	401	409	322	300	328	351
New Users	6	18	10	24	27	218	307	557	1,086	279	211	191	216	211	205
- Referred by Facebook	0	0	0	7	1	205	299	53	1,069	256	198	185	197	203	164

Figure 5. Comparing Unique Pageviews (website traffic) and New Users to the website's resources mapped against CTR traffic coming directly from Facebook.

looking at which terms are yielding the highest CTR. While the second has to be based on feedback from the NGO. The exact terms will not be listed here, as we see large adversarial shifts in this space and coded language evolves quickly. However, in the 3 months after the launch of the Life After Hate redirection in the United States, Facebook saw that when terms were searched related to the redirection list there was on average 4% CTR. While that might not seem very high for those that do not look at marketing data, that is a very competitive CTR compared with brands trying to sell things.

According to third-party analysis across online sites and CTR reviews, the average CTR in AdWords is 1.91% for search related ads surfacing. CTR depends heavily on the industry and product. However, roughly speaking, a good AdWords CTR is 4% to 5%+ on the search network.[35] Our expectations would have been that this had a lower CTR than normal branding and marketing content since the topic is highly sensitive and a person is choosing to enter the domain of an NGO explicitly linked to Facebook flagging that the term that has been input is labeled as "dangerous." The second metric Facebook could look at was which terms were yielding the audience to click through. The top line results around this were that the highest CTR terms were primarily about current groups and individuals that were actively operating in the United States and, to some extent, in the United Kingdom. These were groups and individuals that would have been actively in the media for various reasons, or that might be active within the community in the United States. There was little to no CTR data around broader legacy groups, more obscure slur terms or historical white supremacy branding phrases. The only exception to this was around terms relating to Hitler. These are positive findings in and of themselves though it is hard to test fully for false positive audiences and we cannot fully measure any sentiment shift in the audience without comparing these results to results from the NGO.

Life After Hate offered to share some of their basic site web analytics and practitioner insights data that they had collected. The redirect initiative was launched in March 2019. Life After Hate began partitioning out basic analytics in December 2018 to identify any correlating or significant increases or variations in the data. Their results added to this analysis reflect measurement between December 2018 and February 2020 and include measuring: new users and referrals from Facebook, unique page views and inquiries.

The above table in Figure 5 maps Unique Pageviews and New Users to the platform in comparison to CTR traffic from Facebook. While the below looks at new users and referrals from Facebook more closely in a graph form. Original expectations were that there would be an initial increase from any news coverage and Facebook announcements around the launch of the partnership, but we might expect a normalizing and

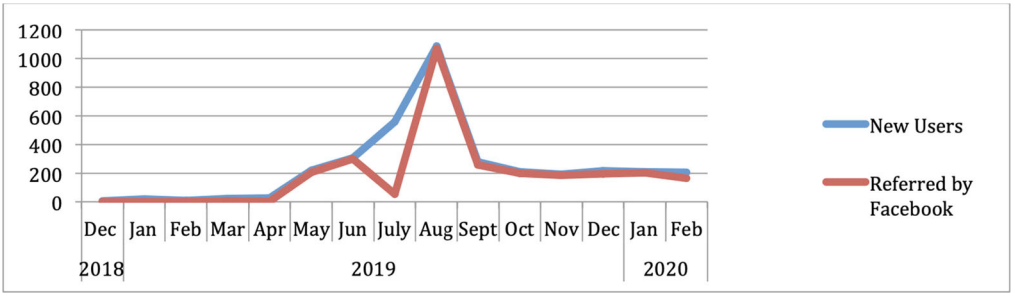

Figure 6. New users and referrals from Facebook (landing on lifeafterhate.org/media): 1 December 2018–29 February 2020. The above chart shows new users and referrals from Facebook in a graph form, compared to the previous table depiction in Figure 5.

some relevant consistent increase in the months following the launch. All the metrics show a distinct and consistent increase even 12 months after the initial launch.

Looking at new users and referrals from Facebook it is clear that the NGO went from nearly zero referrals based on Facebook resources and click throughs to average out over time to ±200 new users to their websites per month. This 200% increase is the most impactful metric as it speaks not just to the passive traffic that a CTR might lead to a webpage, but active engagement with disengagement material and practitioners. This initiative correlates with timing of a few dozen more official and longer-term engagements with individuals looking for assistance in leaving violent extremist organizations and movements. This reflects, anecdotally, a Redirect Initiative taking a passive search function that results in an active engagement as well as individuals choosing to further engage in a process to disengage from a hate-based and/or violent extremist ideology. This is potentially a positive indicator of initiative and program efficacy and begins to get at sentiment change.

While ±200 individuals a month might not seem like a large number comparatively with the large population of the United States, this audience is the hardest audience to reach. They are the low-prevalence/high-risk audience type that is often most difficult to reach. Defined by their active queries and searches for labeled violent extremist terms, groups and individuals, the ability to provide an alternative search result that can facilitate their disengagement in practical terms is crucial. While false positives cannot be wholly ruled out, we assume that the majority of individuals querying for these specific terms and phrases, and then clicking through to the module, belong to the low-prevalence/high-risk audience type because of the very specific terms chosen for the redirection. While some interested parties in the CTR data might include journalists and academics researching around these groups and movements, for our NGO partner to have sustained click through and website engagement, as well as outreach, goes beyond the limited false/positive potentials.

The metrics all point to positive increases in active engagement with Life After Hate resources, as visualised in Figure 6 showing the increase in new user referrals from Facebook to the Life After Hate site, which can be causally attributed in large part to the Facebook redirect initiative. That being said, there is one aspect apparent in the above metrics that we have not accounted for: the large increase in new users and referrals from late July 2019 to early September 2019. It is assumed that this was triggered

by the influence of real-world events and news cycles around the extreme right, white supremacy and neo-Nazi trends in America. However, neither Facebook nor the NGO were tracking the offline, real-world factors that led to this spike at the time, limiting our ability to fully analyze causality.

This pilot project is promising and may be indicative of a positive outcome and initial metrics showing the ability to turn passive online consumption into active disengagement from violent extremist ideologies through the Redirect Initiative approach. Based on the initial success of this program, the Redirect Initiative has now launched in Australia, Indonesia and Germany with local, in-country NGO partners.

What also is clear from this initial pilot is that third party analysis would be helpful to mitigate the information gap between Facebook and the NGO partners it works with. Having a third-party help assess the disengagement process more thoroughly and assist in analyzing the user experience on the NGO website is something that Facebook should not be doing, given its role and limitations around privacy concerns and respect for the NGO partner. To fill this gap and further the evaluation of the Redirect Initiative, Facebook has partners with Moonshot CVE, a U.K.-based counter-extremism NGO with a data-driven approach, and international experience working with tech companies, former extremist programs and disengagement modules. They were also an initial consultant on the original Redirect Method. As this program grows, having a third-party expert NGO to develop, refine and evolve the progress of these efforts internationally and methodically consider measurement and evaluation will be crucial.

Conclusions and Next Steps

Based on the findings from testing both models, this research makes the following six conclusions.

1. **Exposure to counterspeech does not show unintentionally negative effects on the audiences that counterspeech content is surfaced to.** Some articles in recent years have speculated that there are potential negative effects surrounding exposure to counterspeech, even going so far as to say that deployment of counterspeech as a method might not be justifiable.[36] Looking at the two methodologies deployed, it was important to measure whether or not any indicators around negative effects to exposure of counterspeech could be seen. In particular, with the A/B Testing, we proactively looked at the data to see whether or not there was any evidence of further radicalization, or further engagement/viewership of violent extremist content/networks after exposure to the counterspeech treatment series. There is no data that was found to support the theory that exposure to counterspeech negatively impacts behavior or sentiment indicators. There was no data showing "increased radicalization" based on either test.
2. **Indications of positive behavioral change could be assessed in the more refined, low-prevalence high-risk audience types in the aftermath of exposure to a series of counterspeech**. To be accurate, based on the A/B test model, behavioral change in the wider target audience was not statistically significant overall. It was net neutral showing no significant behavioral difference with the

at-risk audience that had not received the counterspeech treatment. However, when the audience was further segmented to account for remaining false/positive potentials, among the higher risk target audience, positive shifts of decreased engagement with violent extremist content was notable. This refined at-risk audience decreased and, in some cases, stopped sharing further violating content and the type of content shared on Facebook was neutralized. The need to further refine the target audience leads us to an important secondary takeaway.

3. **To ensure that counterspeech is reaching the right target-audience, a combination of hard behavioral indicators needs to be deployed in defining the audience.** Most counterspeech to date has been deployed through ads marketing tools, which will always be limited in ensuring that content is reaching its intended target audience. For that reason, measuring and evaluating the success of targeted counterspeech that relies solely on ads marketing to reach at-risk audiences will be limited. Focusing on behavioral signals to define an at-risk audience more accurately ensures that campaigns deployed are reaching the intended audience, thereby cutting out the noise around results and ensuring cleaner data for analysis. This includes triggers based on an individual sharing, engaging or searching for known violating extremist content, groups and/or individuals.

   To note: the broader the aim and target audience the counterspeech is intended for, the more likely ads marketing tools can be effective. For example, if the aim of the counterspeech is to be preventative, to reach broader groups of "young people" as a target audience, then the age, likes and interest inputs in ads tooling can be effective. The more that counterspeech aims to engage "downstream" audiences that are actually "at-risk" of a certain form of radicalization due to initial curiosity and/or sympathies toward that violent extremist ideology, the less effective broad ads-marketing tools will be.

4. **A single hard indicator is often not enough to accurately define an individual as "at-risk."** One-off signals of a shared piece of violent extremist content, or the act of searching out information around dangerous groups, are not clean indicators for an individual's sympathies or membership in a violent extremist group. As found in the A/B test, despite using a hard indicator such as sharing a piece of violent extremist content, as the trigger to include an individual in the counterspeech trial results had to be further refined, as mentioned above, to look deeper at audiences that we could assess were actually displaying potential initial sympathies to a violent extremist ideology. In the qualitative assessment of the A/B test, those that had shared one-off piece of violating content, on further analysis, could be labeled as journalists, academics or individuals vocalizing sentiments against the violent extremist group.[37]

5. **Counterspeech content in video form should aim to be as concise and short in timing as possible with the main takeaway messages, or call to action, presented clearly as early as possible.** In the A/B test model, retention on videos averaged between a few seconds up to a minute or two, whereby most of the counterspeech put forward by NGOs was significantly longer. English language content surfaced in the A/B test averaged 6.01 min and for the Arabic language

content it averaged 3.36 min. If the main messaging, aim and/or call to action of the counterspeech is left toward the end of content there is possibility that the main goal of the counterspeech is lost.

6. **Passive content searches leading to clear options for active counterspeech engagement has the greatest potential for measuring positive sentiment shifts**. Based on the Redirect Initiative, we have initial results of exponential increases in audiences finding online resources leading them to off-platform engagement with disengagement practitioners and resources. This unidirectional, search-based approach to counterspeech has yielded some of the most promising initial results in an online engagement reaching the hardest to reach audiences, leading to human resourced help and continual engagement. More research into these initial findings is needed since the metrics and measurement lies between both Facebook and third-party NGOs.

Next Steps

Each methodology not only showed initial positive results but also other areas to further develop and scope for expansion into different use cases. There is a need to incorporate the nuances of different geographies and counterspeech addressing different forms of violent extremist ideologies. The A/B testing methodology focused on counterspeech aimed at undermining Islamist extremist sympathies for audiences based in the U.K. and Iraq. In comparison, the Redirect Initiative focused on connecting white supremacy and neo-Nazi sympathizers with active disengagement practitioners in the United States. In both cases, a diversity of geographic and ideological aims would help broaden understandings around effectiveness. Since the initial results of the U.S. redirection launch, Facebook has subsequently launched similar redirection modules with NGO partnerships in Australia, Germany and Indonesia.

Both methodologies are applicable for other safety and risk mitigation areas. While the Redirect Initiative built off of methodologies originally deployed for mitigating drug abuse, among other harm areas, there is great scope to expand A/B Testing to other countries and use cases. This is particularly relevant to expanding ongoing safety work around child safety and risk mitigation of human exploitation where Facebook already has a range of counterspeech Ads campaign NGO partners. These methodologies are also potentially relevant to partnerships built to react to real-world crises, indicators around suicide and self-injury, sextortion, as well as mental health help resource surfacing. In many cases, the credible partnerships between localized NGOs and the private sector already exist in various programmatic forms and could be optimized by the utilization of these two methodologies.

Best practices can be better conveyed through partnerships between the private sector collaborating with civil society, NGOs and experts. Both methodologies and many other risk mitigations approaches online rely on partnerships with credible, localized practitioners and activists. As discussed in the introduction to this article, Facebook and other private tech companies are not the credible voice nor the ground-level practitioner to strategically counter-extremism. However, tech companies are the tool and infrastructure that can help upscale and optimize the connection between the practitioners and

their resources in reaching refined at-risk audiences. Both methodologies were reliant on nuanced partnerships with NGOs who either develop localized counterspeech content or who were disengagement practitioners. In the cases where evaluation metrics rely on data from a third-party NGO, it is beneficial to partner with experts who can facilitate evaluation and look to co-develop best practices. This is currently taking place in a partnership between Facebook and Moonshot CVE in order to evaluate and facilitate the scaling up of the Redirect Initiative.

Notes

1. Google and Microsoft have also developed proactive models to surface counterspeech through the methodology first developed by Jigsaw, called the Redirect Method. The Redirect Method is meant to work in a manner whereby when online users of a given search engine enter certain violent extremist related search terms into YouTube, Google Search or Bing Search they are presented with a mixture of curated counterspeech content and a counterspeech playlist of videos. This methodology was developed around combatting Islamist extremist recruitment to the so-called Islamic State. See "Redirect Method," accessed February 2, 2020, https://redirectmethod.org.
2. *Literature Review of Terrorist and Violent Extremist Use of the Internet,* the research reviewed 574 articles on the topic of terrorist and violent extremist use of the internet by the University of Queensland. This research showed that the largest two areas of extremism with an online nexus concern "Right-Wing Extremism" and "Islamic extremism." It also pointed out that there is often a lack of comparative exploration of different online environments with regards to language and geography, the role of underlying technologies and a lack of metrics speaking to the efficacy of CVE measures. See: Winnifred Louis, Susilo Wibisono, Kevin M Blasiak, and Marten Risius, *Literature Review of Terrorist and Violent Extremist Use of the Internet* (Australia: The University of Queensland, October 2020).
3. See programs discussed on the Facebook online Counterspeech hub: Facebook Counterspeech, Facebook. https://counterspeech.fb.com/en/ (accessed March 3, 2020).
4. Terrorists and violent extremists have not only always relied on a variety of communication technologies to connect their in-group with its members but also more broadly to relay communications to potential recruits as well as enemies and wider public. These networks have taken advantage of the internet as a medium to communicate with mass audience and smaller in-groups. See: Anne Aly, Tom Chen, Lee Jarvis, and Stuart Macdonald, "Introduction," in *Violent Extremism Online: New Perspectives on Terrorism and the Internet*, ed. Anne Aly, Stuart Macdonald, Lee Jarvis, and Thomas Chen (New York: Routledge, 2016), 1–7.
5. See: Michael Kenney, "Beyond the Internet: Mētis, Techne, and the Limitations of Online Artifacts for Islamist Terrorists," *Terrorism and Political Violence 22*, no.2 (2010): 177–97. See also: Miron Lakomy, "Let's play a video game: Jihadi propaganda in the world of electronic entertainment," *Studies in Conflict & Terrorism* 42, no.4 (2019): 383–406. See also: Sean C. Reynolds and Mohammed M. Hafez, "Social Network Analysis of German Foreign Fighters in Syria and Iraq," *Terrorism and Political Violence* 31, no.4 (2017): 661–86. See also: Diana Rieger, Lena Frischlich, and Gary Bente, "Dealing with the Dark Side: The Effects of Right-Wing Extremist and Islamist Extremist Propaganda from a Social Identity Perspective," *Media, War & Conflict* 13, no.3 (2019): 280–99.
6. See: Ahmed Al-Rawi, "Video Games, Terrorism, and ISIS's Jihad 3.0," *Terrorism and Political Violence* 30, no.4 (2018): 740–60. See also: Lars Guenther, Georg Ruhrmann, Jenny Bischoff, Tessa Penzel, and Antonia Weber, "Strategic Framing and Social Media Engagement: Analyzing Memes Posted by the German Identitarian Movement on Facebook," *Social Media + Society* 6, no.1 (2020): 1–13. See also: Lauren R. Shapiro and

Marie-Helen Maras, "Women's Radicalization to Religious Terrorism: An Examination of ISIS Cases in the United States." *Studies in Conflict & Terrorism* 42, no. 1–2. (2019): 88–119.

7. Anne Aly, Stuart Macdonald, Lee Jarvis, and Thomas Chen, *Violent Extremism Online: New Perspectives on Terrorism and the Internet* (New York: Routledge, 2016).
8. See: Cristina Archetti, *Understanding Terrorism in the Age of Global Media: A Communication Approach* (London: Palgrave Macmillan, 2012). See also: Paul Cornish, Julian Lindley-French, and Claire York, *Strategic Communications and National Strategy* (Chatham House, The Royal Institute of International Affairs, 2011). See also: Tim Stevens and Peter Neumann, *Countering Online Radicalisation: A Strategy for Action*, (London: The International Centre for the Study of Radicalisation and Political Violence (ICSR), 2009). https://cst.org.uk/docs/countering_online_radicalisation1.pdf.
9. See: Carol K. Winkler and Cori E. Dauber, *Visual Propaganda and Extremism in the Online Environment,* (Strategic Studies Institute and U.S. Army War College Press, 2014). For more on the necessary messenger and charisma see: Angela Gendron, "The Call to Jihad: Charismatic Leaders and the Internet," in *Violent Extremism Online: New Perspectives on Terrorism and the Internet*, ed. Anne Aly, Stuart Macdonald, Lee Jarvis, and Thomas Chen (New York: Routledge, 2016), 25–44. For more on the packaging of jihadist messages using visual motifs, locations and hymns see: Arab Salem, Edna Reid, and Hsinchun Chen, *The Islamic Imagery Project: Visual Motifs in Jihadi Internet Propaganda* (West Point, New York, USA: Combating Terrorism Center, 2006).
10. Interviews were held with leading researchers, experts and practitioners from; ISD Global, Moonshot CVE, Quilliam Foundation, ConnectFutures, Groundswell, and individual members from the Against Violent Extremism Network.
11. For an overview on Facebook's global counterspeech programmatic efforts, resources and partnerships see Facebook's Counterspeech page, Facebook. https://counterspeech.fb.com/en/ (accessed March 3, 2020).
12. The need to localize and better connect the right content with the intended target audience is discussed at length by a range of researchers. See: Anne Aly, "Brothers, Believers, Brave Mujahideen," in *Violent Extremism Online: New Perspectives on Terrorism and the Internet*, ed. Anne Aly, Stuart Macdonald, Lee Jarvis, and Thomas Chen (New York: Routledge, 2016). See also: Cori E. Dauber and Carol K. Winkler, "Radical Visual Propaganda in the Online Environment: An Introduction," in *Visual Propaganda and Extremism in the Online Environment,* ed. Carol K. Winkler and Cori E. Dauber (Strategic Studies Institute and U.S. Army War College Press, 2014), 1–30. See also, Lars Guenther, Georg Ruhrmann, Jenny Bischoff, Tessa Penzel, and Antonia Weber, "Strategic Framing and Social Media Engagement: Analyzing Memes Posted by the German Identitarian Movement on Facebook," *Social Media + Society* 6, no.1 (2020).
13. See: Anne Aly, Stuart Macdonald, Lee Jarvis, and Thomas Chen, *Violent Extremism Online: New Perspectives on Terrorism and the Internet* (New York: Routledge, 2016). See also: Angela Gendron, "The Call to Jihad: Charismatic Leaders and the Internet," in *Violent Extremism Online: New Perspectives on Terrorism and the Internet*, ed. Anne Aly, Stuart Macdonald, Lee Jarvis, and Thomas Chen (New York: Routledge, 2016): 25–44.
14. There is ongoing discussion about the need for greater transparency across social media with respect to how it counters terrorism and violent extremism. With reference to greater transparency in counterspeech and counter-extremism efforts see: Jamie Bartlett and Carl Miller, *The Power of Unreason: Conspiracy Theories, Extremism and Counter-Terrorism* (London: Demos, 2010).
15. The outline shows where in a process an intervention might be (1) prevention of violent extremism, (2) countering violent extremism or (3) countering terrorism is based off of source interviews with practitioners described within the methodology of this article. This model expands upon the pyramid of the *Three Levels of Prevention*, as documented in: Erin Saltman and Melanie Smith, *Till Martyrdom Do Us Part: Gender and the ISIS Phenomenon* (London: ISD Global, 2016): 53, https://www.isdglobal.org/wp-content/uploads/2016/02/Till_

Martyrdom_Do_Us_Part_Gender_and_the_ISIS_Phenomenon.pdf. This model also builds off of the Staircase Model of Radicalization in: Fathali M Moghaddam, "The Staircase to Terrorism: A Psychological Exploration," in *American Psychologist* 60, no.2 (September 2007): 161–69.

16. Ron Kohavi and Roger Longbotham, "Online Controlled Experiments and A/B Testing," *Encyclopedia of Machine Learning and Data Mining* 7, no.8 (2017): 922–9.
17. Bulpitt Ci and Round-Off Rule, "Confidence Intervals," *Lancet* 1 no. 8531 (1987): 494–7.
18. All content identified as terrorist or hate-based organization content, once surfaced or flagged, is subsequently removed for violating Facebook's Dangerous Organizations Policy. See: "Dangerous Individuals and Organizations," Facebook Community Standards. https://www.facebook.com/communitystandards/dangerous_individuals_organizations (accessed January 17, 2020).
19. For more about ConnectFutures visit, <www.connectfutures.org>.
20. Jamie Bartlett and Alex Krasodomski-Jones, *Counterspeech: Examining Content that Challenges Extremism Online* (London: Demos, 2015). https://www.demos.co.uk/wp-content/uploads/2015/10/Counter-speech.pdf (accessed September 1, 2020).
21. See: Louis Reynolds and Henry Tuck, *The Counter-Narrative Monitoring and Evaluation Handbook*, (London: ISD Global, 2016). https://www.isdglobal.org/isd-publications/the-counter-narrative-monitoring-evaluation-handbook/ (accessed May 20, 2020). See also: *Online Civil Courage Initiative (OCCI): Information Pack on Counterspeech Engagement* (Facebook & ISD Global, 2017). https://counterspeech.fb.com/en/wp-content/uploads/sites/2/2017/06/occi-counterspeech-information-pack-english.pdf (accessed June 20, 2020).
22. Leo Kirby, Margaret Williams, and Jorge Ramirez Mata, "Counterspeech: How Does Online Counterspeech Impact Real World Behavior?," IPS Practicum MA Project, (Stanford University, 2018).
23. For more about Facebook's Data Policy see < facebook.com/policy.php>.
24. After reviewing our testing data, we determined that dividing the candidates into three distinct audience types based on earlier content violations would allow us to provide an objective description of user behavior. In the UK, 77% did not have an earlier violation (i.e., did not have any of their content taken down for terrorism reasons), but have engaged with violating content. Six percent had one violation, and the remaining 17% had two or three earlier violations. In Iraq, 68% did not have an earlier violation, 19% had one, and 13% had two or three earlier violations.
25. "Standing Against Hate," Facebook Newsroom, Facebook, 27 March 2019. https://about.fb.com/news/2019/03/standing-against-hate/ (accessed May 12, 2019).
26. Researchers such as Michael Kenney (2010), Miron Lakomy (2019), Sean Reynolds and Mohammed Hafez (2017), and Diana Rieger, Lena Frischlich, and Gary Bente (2019) have argued that passive consumption of terrorist or violent extremist content plays less of an overall role in the process radicalization. Interpersonal connections and community remain crucial though some of the interpersonal connectivity can happen online. See: Michael Kenney, "Beyond the Internet: Mētis, Techne, and the Limitations of Online Artifacts for Islamist Terrorists," *Terrorism and Political Violence 22*, no.2 (2010): 177–97. See also: Miron Lakomy, "Let's Play a Video Game: Jihadi Propaganda in the World of Electronic Entertainment," *Studies in Conflict & Terrorism* 42, no.4 (2019): 383–406. See also: Sean C. Reynolds and Mohammed M. Hafez, "Social Network Analysis of German Foreign Fighters in Syria and Iraq," *Terrorism and Political Violence* 31, no.4 (2017): 661–86. See also: Diana Rieger, Lena Frischlich, and Gary Bente, "Dealing with the Dark Side: The Effects of Right-Wing Extremist and Islamist Extremist Propaganda from a Social Identity Perspective," *Media, War & Conflict* 13, no.3 (2019): 280–99.
27. Others have argued that violent extremist content can allure and radicalize viewers toward violence, highlighted in research by Ahmed Al-Rawi (2018), Lars Guenther, Georg Ruhrmann, Jenny Bischoff, Tessa Penzel and Antonia Weber (2020), and Lauren Shapiro and Marie-Helen Maras (2019). See: Ahmed Al-Rawi, "Video Games, Terrorism, and ISIS's Jihad 3.0," *Terrorism and Political Violence* 30, no.4 (2018): 740–60. See also: Lars Guenther,

Georg Ruhrmann, Jenny Bischoff, Tessa Penzel, and Antonia Weber, "Strategic Framing and Social Media Engagement: Analyzing Memes Posted by the German Identitarian Movement on Facebook," *Social Media + Society* 6, no.1 (2020). See also: Lauren R. Shapiro and Marie-Helen Maras, "Women's Radicalization to Religious Terrorism: An Examination of ISIS Cases in the United States." *Studies in Conflict & Terrorism* 42, no.1–2. (2019): 88–119.

28. "Safety Center," Facebook. https://www.facebook.com/safety (accessed March 4, 2020).
29. Greg Williams, "How Facebook is Addressing the Opioid Crisis," *About Facebook*. https://about.fb.com/wp-content/uploads/2018/11/addressing-the-opioid-crisis-1.pdf (accessed March 4, 2020).
30. "Redirect Method." https://redirectmethod.org (accessed February 2, 2020).
31. Find out more about Life After Hate on: Life After Hate. https://www.lifeafterhate.org/ (accessed May 4, 2020).
32. Find out more about Exit USA on their landing page: Exit USA. https://www.lifeafterhate.org/exitusa (accessed May 4, 2020).
33. More about click through rates (CTR) and their metrics can be found on the Facebook Business Help Center here. See: "CTR (Link Click-Through Rate)," Facebook Business Help Center, Facebook. https://www.facebook.com/business/help/877711998984611 (accessed November 1, 2020).
34. John E Richardson, *A Guide to Online Radical-Right Symbols, Slogans and Slurs* (Oxford: Centre for Analysis of the Radical Right, April 2020). http://www.radicalrightanalysis.com/2020/05/04/a-guide-to-online-radical-right-symbols-slogans-and-slurs/ (accessed June 27, 2020).
35. "Click-Through Rate (CTR): Understanding Click-Through Rate for PPC," *WordStream*. https://www.wordstream.com/click-through-rate (accessed May 12, 2020).
36. Ann-Sophie Hemmingsen and Karin Ingrid Castro, *The Trouble with Counter-Narratives* (Copenhagen: DIIS Reports, Danish Institute for International Studies, 2017). https://www.econstor.eu/handle/10419/197640 (accessed December 2, 2020).
37. This was echoed in the case study of shared content around the New Zealand, Christchurch attacker when Twitter found that violating content of the attacker's footage was shared widely and largely through verified channels, not by sympathizers or those intending to praise, support or glorify the attack. See: Sanjana Hattotuwa, "Pulse Points," *Scoop Independent News*, March 21, 2019. https://www.scoop.co.nz/stories/HL1903/S00124/pulse-points.htm (accessed May 1, 2020).

Acknowledgments

The authors would like to give thanks and acknowledge the NGOs that partnered with Facebook to be the substance of the content launched and surfaced via these two methodologies: Life After Hate, the International Center for the Study of Violent Extremism (ICSVE), the Adyan Institute and ConnectFutures. Thanks also goes to the MA Students at Stanford University – Leo Kirby, Margaret Williams and Jorge Ramirez Mata – who piloted the rate and frequency testing which built a framework for the A/B testing. Special thanks also go to the myriad of Facebook cross-functional teams that donated time and energy toward developing, launching and evaluating these methods across policy, safety, engineering, content, product and integrity teams. This includes, but is not limited to, Brian Fishman, Dina Hussein, Sarah Pollack, Ignacio Contreras, Lauren Ruth Vilders, Sarah Vieweg, Akum Shergill, Zoe Stoll, Igor Shilov, Elizabeth Murphy, Tommy Giglio, Natalie Shaw and Nick Inzucchi. Your dedication to countering violent extremism in all its forms is truly appreciated.

Disclosure statement

No potential conflict of interest was reported by the author(s).