From passive search to active conversation:

An evaluation of the Facebook Redirect Programme

Overview of the pilot programme and recommendations for future deployments

November 2020

Contents

	Executive summary	2
I.	Introduction	4
II.	How the Facebook Redirect Programme was designed	6
III.	Evaluating the Facebook Redirect Programme	8
IV.	Increasing the number of active conversations: observations and recommendations	16
V.	Conclusion and next steps	20
Appendix I: Onboarding Checklist		21
Appendix II: Methodology		23

Executive summary

This report presents Moonshot's evaluation of the Facebook Redirect Programme (FRP), including its design and implementation, along with our findings and recommendations.

The fundamental purpose of the FRP is to ensure that Facebook users searching for dangerous individuals and organisations on the platform are offered authentic, meaningful and impactful support off-platform. The purpose of the pilot was to test the programme design with a view to informing and improving future deployments.

The core components of the FRP are: a list of keywords related to dangerous individuals and organisations; a 'safety module' comprising explanatory text and a hyperlinked call to action; and a delivery partner landing page where visitors can access support services. Facebook and the delivery partners agree on the keyword list; Facebook govern the safety module; and delivery partners govern their own websites and service offerings.

Any Facebook user searching for one of the predetermined keywords using the platform's search function will see the safety module at the top of the results page. It reads, "These keywords may be associated with dangerous groups and individuals. Facebook works with organizations that help prevent the spread of hate and violent extremism. [Learn More]." Clicking "Learn More" redirects users to the website of the delivery partner most relevant to their geographic location.

The programme is currently live in the United States, Australia, Indonesia and Germany, encouraging Facebook users who search for dangerous organisations and individuals to visit the websites of local intervention providers and start a conversation.

The pilot FRP evaluated in this report ran in the United States and Australia between May 2019 and March 2020, redirecting users in the United States to Life After Hate, and in Australia, to Exit Australia. For this evaluation, Moonshot interviewed staff at Facebook and both delivery partner organisations and had full access to the Google Analytics dashboards of both delivery partners, as well as relevant data from Facebook.

Our conclusion is that the pilot of the FRP was broadly successful. This conclusion is based primarily on the fact that, during the three-month pilot, 25 Facebook users who initially sought to engage with violent extremism on the platform instead received some form of support from one of the delivery partners. This demonstrates that the programme is able to successfully link high-risk individuals with support, and help turn a passive search for violent extremist content into an active conversation.

Upstream metrics complement this finding. Tens of thousands of Facebook users were offered help by the safety module, of which thousands accepted and went on to engage with supportive content on delivery partner sites. Of these, 25 individuals chose to begin a conversation. On this basis, our report concludes the following:

- Facebook can use its search module to reach 'low-prevalence, high-risk' audiences. In doing so, they have successfully
    - created friction between the search for a white supremacist and/or neo-Nazi community and a positive result, and to some extent
    - functioned as the conduit between high-risk individuals and their respective delivery partners, which has helped to turn some passive searches into active conversations.

Our evaluation also identified a number of weaknesses in both the design and the implementation of the pilot FRP. These were communicated to Facebook either immediately and/or as part of our internal Pilot Evaluation Report, delivered in August 2020. These included:

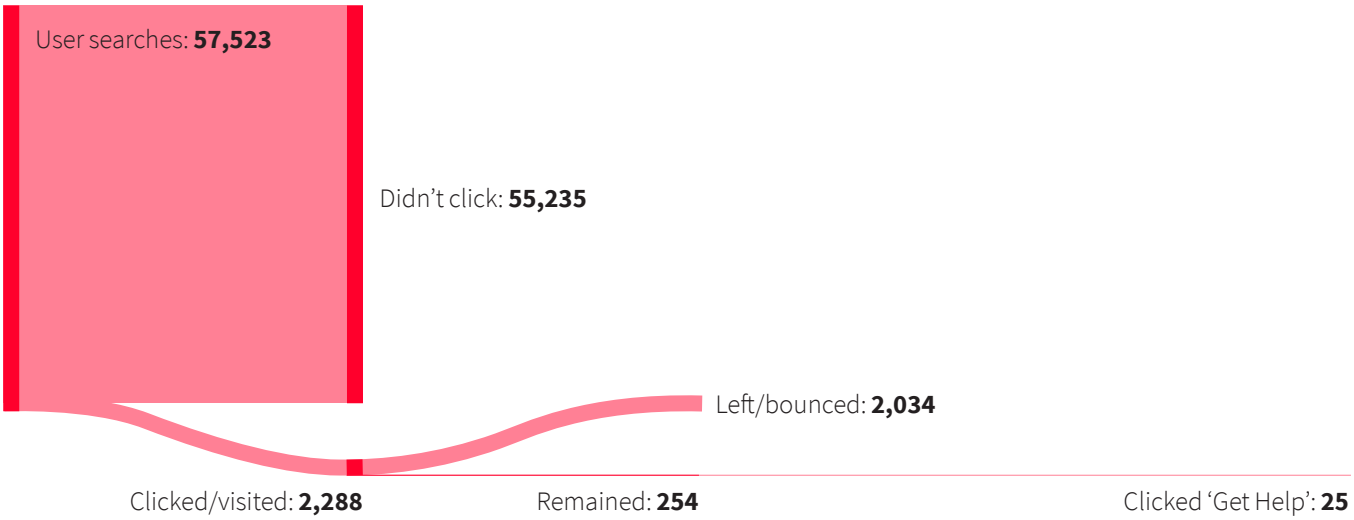
- The absence of a standardised onboarding process for delivery partners;
- The fact the FRP's safety module was appearing in the UK when it was only supposed to be live in the USA and Australia;
- Insufficient staff time dedicated to the programme within Facebook;
- Potentially misleading language in the safety module call to action.

Some of these, along with other criticisms and observations, were specific to the pilot programme and have already been addressed. The error causing the safety module to appear in the UK has been fixed. Facebook now have staff dedicated to the programme and the language of the safety module's call to action has been changed from 'Learn More' to 'Learn More from [delivery partner name]', to make it clear that users will be taken off-platform. While this Final Report makes reference to these findings, the primary focus is on observations and recommendations more applicable to future deployments - chiefly, the need for a standardised process for onboarding delivery partners. To this end, we have drawn up a sample 'Onboarding Checklist' designed for use in future deployments (see Appendix I).

This is an ambitious programme. Any attempt to redirect an individual away from their intended destination and towards one which is not only not where they wanted to end up, but also seeks to challenge their worldview, necessarily faces a series of technical, ethical and user experience challenges. These are especially acute when those individuals are at risk - whether from bullying, harassment or self-harm, or because they are vulnerable, as in the case of the FRP, to the narratives, myths and recruitment tactics of violent extremist groups and communities.

This ambition is also necessary, now more than ever, to protect online spaces and the global communities that use them. We look forward to the continued evolution of Facebook's Redirect Programme.

Facebook Redirect Programme, pilot period of performance (November 2019 - March 2020): user journey flow from passive high-risk search on Facebook to active conversations with intervention providers.



I. Introduction

Facebook launched the pilot of the Redirect Programme with delivery partners Life After Hate in May 2019 and Exit Australia in September 2019. The programme aimed to ensure that individuals searching for white supremacist and/or neo-Nazi communities on Facebook would be offered authentic, meaningful and impactful support off-platform. The purpose of the pilot was to test the programme design and inform future deployments targeting both new geographies and different hate-based communities.

Moonshot was contracted by Facebook to evaluate the pilot period of programme performance and make recommendations for future deployments. For this, we produced two reports: an internal Pilot Evaluation Report in August 2020 and this Final Report in November 2020. For the purposes of both reports, the pilot period of performance is considered to have ended in March 2020, having begun at different times for the two different delivery partners. Our qualitative evaluation runs from May 2019 until March 2020, whereas our quantitative evaluation is confined to November 2019 to March 2020 to allow us to combine delivery partner metrics for a more accurate picture.

Interviews with staff Facebook produced three fundamental questions the pilot needed to answer:

1. Can Facebook use its search module to reach the 'low-prevalence, high-risk' audiences that advertising services are not designed to reach? In doing so, can they successfully:
   a. create friction between search queries for violent extremist and/or hate-based communities and positive search results, at a minimum; and ideally
   b. function as the conduit between high-risk individuals and delivery partners, and help turn a passive search into an active conversation?
2. In the pilot locations of the United States and Australia, had Facebook chosen the right delivery partners? Do high-risk Facebook users engage with those partners and, by extension, appear to consider them credible, legitimate facilitators of this work?
3. Can this pilot programme achieve sufficient credibility at Facebook such that future deployments are made logistically as well as technically possible?

This report evaluates the pilot programme by examining:

1. Facebook's use of keywords and the safety module as a method of redirecting people off-platform;
2. The full user journey from Facebook to delivery partner landing pages;
3. The extent to which the pilot can be considered a proof of concept for future deployments.

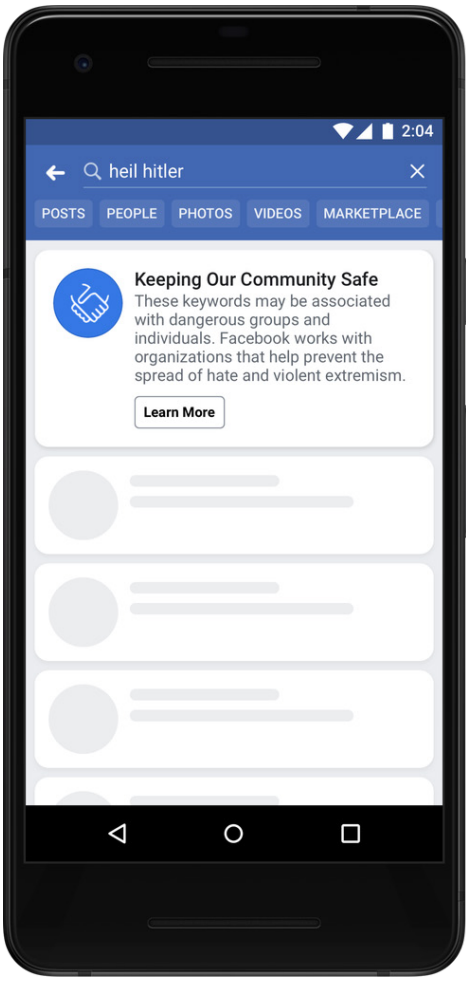
The evaluation is based on semi-structured interviews with staff from Facebook, Life After Hate and Exit Australia, and relevant datasets provided by all three organisations. All data, findings, and insights within this report are anonymised or pseudonymised. Moonshot sought and received consent from all participants. All data is stored and processed in compliance with the General Data Protection Regulation (GDPR).

Background to the Facebook Redirect Programme (FRP)

The genesis of the programme was a series of interviews conducted by Facebook in 2017 with CVE practitioners and former extremists to assess Facebook's areas of strength and weakness in relation to tackling the problem of violent extremist content on its platform. From this process emerged a range of potential counter-hate-speech projects, one of which was the FRP.

Facebook identified search terms entered into the Facebook search bar as the necessary 'hard indicators' that a user was at the very least searching for a dangerous individual or hate-based community, if not part of one already. They wanted to pair this hard indication with a soft resource in a way that was authentic and helpful. It was important that, if Facebook was going to intervene in a user's search results, it was not done in a way that was so jarring or disturbing to users that they turned away from the platform altogether.

Facebook wanted these soft resources to be external, and identified organisations with operational capability, authenticity and local credibility, based on their CVE expertise and deep understanding of the local environments in which they were operating. The pilot was first launched with Life After Hate and later with Exit Australia, both of whom focus on neo-Nazi and white supremacist extremism in the United States and Australia respectively.



The Safety Module

During the pilot, this version of the safety module appeared at the top of users' search results when they had searched for white supremacist and/or neo-Nazi individuals or communities. 'Learn More' took users to the website of either Life After Hate or Exit Australia, where they could find supportive content and access interventions services.

II. How the Facebook Redirect Programme was designed

This section identifies the core components of the programme and explains how Facebook approached and made decisions about their design.

To be functional, the FRP requires three core components:

1. A database of keywords;
2. A 'safety module', triggered by use of the keywords, with a call to action hyperlinked to a delivery partner website;
3. A delivery partner website with
   a. The ability for the user to seek assistance or services, e.g. by filling out a form and awaiting a callback.

To be effective, each of these components requires a level of quality:

1. The keywords need to capture the right kind of intent - in this case, a desire to find neo-Nazi and/or white supremacist communities on the platform.
2. The safety module needs to be written in accordance with internal best practice such that it interrupts but does not disturb the expected user journey (where 'disturbance' would lead to the user turning away from the platform).
3. The landing page on the delivery partner website needs to function as a logical continuation of the user journey, and the ability to self-refer to support services needs to be well-placed and easy to use.

Selecting and preparing the delivery partners

Facebook's criteria for delivery partners

1. Independent of government;
2. Able to evidence a history of providing violent extremist disengagement resources or deradicalising services successfully and legitimately;
3. Currently providing the above services via their website;
4. Geographically relevant to the target audience;
5. Interested in partnering with Facebook.

The above criteria for delivery partner selection led the approach to their procurement. Life After Hate was initially approached by Facebook, whereas Exit Australia contacted Facebook having recognised the need for the same work in Australia in the wake of the Christchurch attack in March 2019. Facebook's internal public policy team was tasked with making an official assessment of the local reputation and legitimacy of both prospective organisations.

The preparations and support given to delivery partners was limited, according to interviews with both Facebook and delivery partner staff. Facebook flagged safety issues such as removing personally identifiable staff information from their websites, as well as other digital hygiene best practices. Facebook also offered Life After Hate feedback directly from their marketing team.

Interviews with Facebook staff indicated that this level of support was intentional. They were keen to respect the independence of delivery partners and their processes. In particular, they felt it would be inappropriate to issue guidance on service provision given the delivery partners' expertise.

Choosing the keywords

Keyword selection was a collaborative effort between the individual delivery partners and Facebook. Both delivery partners provided Facebook with an initial long-list of potential terms, which Facebook then ran through an extensive testing and quality assurance process. Multiple teams within Facebook reviewed every keyword to check, among other things: whether it had alternative meanings; how much of the content related to the violent extremist or hate-based network in question; and whether the term had different meanings in different geographies.

Designing the safety module

Searching for one of the predetermined keywords triggered the safety module to appear above the user's search results. It reads: 'Keeping Our Community Safe. These keywords may be associated with dangerous groups and individuals. Facebook works with organisations that help prevent the spread of hate and violent extremism. [Learn More].'

Best practice around the user experience elements of the safety module were based on Facebook's previous work providing soft resources to users searching for content related to a number of harmful topics, including self-harm, bullying and harassment, and the opioid crisis. For this deployment of the module, Facebook relied on this established best-practice rather than, for example, A/B testing different modules. As such, the appearance of the module, including the headline 'Keeping our community safe', the badge and its colour were all carried over from previous deployments.

The language was bespoke to this deployment and Facebook UX researchers were brought in to help craft the message. The aim was to be friendly and neutral, similar to Facebook's usual tone of voice, and to avoid giving the impression users were being followed or accused of something. Similarly, 'Learn more' is a deliberate construction which gives the user an incentive to act, rather than the more didactic, 'Click here'.

Clicking on 'Learn More' took the user to a landing page belonging to either Exit Australia or Life After Hate. To determine which landing page 'Learn More' would redirect to, the Facebook platform considered the location of the user, the language in which the user conducted their search and the search term - the final automated decision being a combination of all three factors.

Redirecting users

During the pilot, users whose searches triggered the safety module for Exit Australia were redirected to a page of videos about both jihadist and violent far-right extremism. From there the individual could click on a link in the navigation bar labelled 'Get Support' in order to contact the organisation via a form, or contact national emergency services via numbers provided on the page.

During the pilot, users whose search triggered the safety module that redirected to Life After Hate were redirected to a page of videos about far-right extremism and extremism more generally. At the top of the page there was a 'Get Help' link, which took the user to a page detailing the help offered by the organisation, alongside a form and separate phone number, both of which could be used to contact them.

III. Evaluating the Facebook Redirect Programme

Project design and process

The objectives of the pilot from Facebook's perspective were to ensure appropriate delivery partner selection, test the smoothness of user journey off-platform, understand the potential impact of the programme based on data, and establish internal and external credibility for the programme. This section evaluates the design of the FRP.

Data provided to Moonshot CVE used in this section

- Semi-structured interviews with staff at Facebook, Life After Hate and Exit Australia.

Selecting and preparing the delivery partners

It is important to note that in evaluating the selection of delivery partners, Moonshot CVE is not evaluating the ultimate decision to partner with Life After Hate or Exit Australia specifically, but rather Facebook's process: that is, the criteria prescribed (and not prescribed) by Facebook that would need to be met by any delivery partner, and the support provided (and not provided) by Facebook to Life After Hate and Exit Australia to enable them to meet those criteria.

The criteria used by Facebook to select delivery partners for the pilot balanced the flexibility necessary for the different contexts with the need for certainty of delivery - an entirely sensible approach given how few non-governmental organisations are available to do this kind of work. Facebook also provided financial support to both delivery partners to cover the costs associated with participation and increased caseload as a result of the programme. This was critical to the ability of delivery partners to be able to deliver the programme safely and effectively.

Choosing the keywords

Having submitted their initial lists, neither delivery partner was entirely clear as to how the keyword selection process would proceed. While final decisions on inclusion and exclusion necessarily sat with Facebook, greater clarity around that decision making would have been beneficial for all parties and potentially for the quality of the lists themselves.

In interviews, Facebook staff were clear that trying to surface the safety module to a low-prevalence, high-risk audience meant that search keywords did not need to have already recorded data on the platform to be included in the programme. Facebook staff also shared with delivery partners a document containing both policy and operational feedback on keywords. Despite this, delivery partners were not always clear as to exactly which keywords were being used and the reasons why others had been removed.

Given the importance of this element of the FRP, future deployments would benefit from increased communication about both the process itself and the decisions made.

Designing the safety module

The safety module was designed in line with best practice from Facebook's previous and ongoing work on providing soft resources to high-risk users across a number of harmful topics, including self-harm, bullying and harassment, and the opioid crisis. Therefore, for this deployment of the module, no testing was carried out on its appearance, including the headline 'Keeping our community safe', the badge and its colour.

It is entirely reasonable to take inspiration from what has worked well in the past or for other topics, especially when tone of voice and user experience is critical and alterations to the safety module require high-level approval. However, Facebook should remain open to testing different language and placement for the specific purpose of increasing engagement with delivery partners among the low-prevalence, high-risk audience seeking to engage with violent extremist content. As will be discussed in more detail in the following section, with the data available to us, we know that (a) the user journey could be improved, (b) the safety module has a critical part to play in the user journey, and yet (c) only one version of the module has been tested in this context. Experimenting with different language, calls to action and placements would enable Facebook (and an audience of CVE practitioners) to see if it is possible to connect a greater percentage of the thousands of users searching for these keywords with the right support.

Redirecting users: onboarding delivery partners and minimising risk

Once Life After Hate and Exit Australia had been selected, there was no standard onboarding approach or guidance. We understand Facebook wished to remain flexible and hands-off, in that delivery partners should be allowed to operate as they normally would, but for future deployments it is critical that important pieces of information, such as when high-risk users are going to start being redirected to their services, are communicated to delivery partners as part of the onboarding process. In interviews, Facebook staff were clear that although there was no set timeline, go-live dates were in fact communicated to both delivery partners in advance. However, delivery partners were equally clear in their interviews that they would have preferred more involvement and clarity from Facebook in the run-up to going live.

More so than any other piece of preparatory information, delivery partner staff would have benefitted from more insight into the possible effects of the programme on their workload. Delivery partners were able to accommodate increases in referrals during the pilot, but as we set out in more detail in our recommendations, it is imperative that the risk of them being unable to do so is minimised such that they are always able to deliver the support promised by the FRP to high-risk users.

The absence of any forecasts or estimates of new referrals can largely be put down to this being a pilot. While it would have been technically possible for Facebook to obtain historical search data on the agreed lists of keywords, no such historical data was available for click-through rates, because the safety module had never been used in this context before. Even the lessons learned from this pilot may be of limited use in future deployments, as propensity to click on the safety module's call to action will vary by culture, geography and subject matter, and fluctuate in response to external events.

Nonetheless, it remains imperative to minimise to the greatest extent possible the risk of high-risk individuals not receiving the support promised to them by the FRP. To that end, this pilot does afford the opportunity to create case studies detailing the experiences and lessons learned by the two delivery partners. In particular these should focus on unforeseen increases in caseloads; adjustments to landing page content and language they would have made in hindsight; insights into user journeys derived from Google Analytics; and any other observations that could be useful to newly onboarded, future delivery partners.

In addition and acknowledging the caveats above, Facebook should aim to provide delivery partners with historical data on keyword search volumes prior to their go-live date. The usefulness of this data will vary according to circumstance, but there will be times when cultural comparisons and sensible assumptions about click-through rate data can be made, to the extent that delivery partner preparedness is maximised and the risks are minimised. This data would also inform the levels of financial support required from Facebook to ensure any changes in caseload are adequately resourced.

Redirecting users: optimising landing pages for high-risk individuals

Facebook gave some advice to both delivery partners on how best to adjust their landing pages. This was anticipated and welcomed by both Life After Hate and Exit Australia, but both said they would have liked more input. We appreciate and support Facebook's position that local delivery partners know their audiences best, but feel there is scope in future deployments for delivery partners to receive more support in optimising what is ultimately a critical component of the FRP.

> *"We would be interested in getting support on tone of copy and making sure the language we use is not judgemental. It would also be great to get some help tracking attribution, setting goals and increasing conversions."* [1]

Both delivery partners have the expertise and personnel to make regular adjustments to the design and content of their sites in response to new data. Since the pilot period, both organisations have taken it upon themselves to make changes to their websites with the aim of making their content more relevant to their new Facebook audience.

Evaluation: Project implementation

This section evaluates how the choices made in the design phase performed during the project's implementation.

Moonshot evaluated project implementation based on:

1. the outcomes of the activities in our logic model;
2. interviews with key staff at Facebook, Exit Australia and Life After Hate; and
3. Google Analytics data, keyword lists and click-through rate data.

Data provided to Moonshot CVE and used for the purposes of pilot evaluation

- Semi-structured interviews with staff at Facebook, Life After Hate and Exit Australia
- Full list of keywords which, when entered into Facebook's search bar, trigger the safety module.
- Click-through rate data from Facebook.
- Google Analytics data from delivery partner websites.

[1] Quote from interviews with Life After Hate staff.

User journey, stage 1: Were the keywords the right keywords?

Keyword selection is a critical component of the FRP. Any pre-criminal CVE or CT programming designed to capture user intent by drawing up a predetermined list of keywords - which users may or may not actually use in practice - will face the challenge of balancing breadth and specificity. Too broad, and the high-risk users will disappear into a crowd of false positives; too specific, and some high-risk users will be missed altogether.

Facebook faced both this challenge and an additional problem. They were deciding to interrupt their own users' journeys by telling them that the term they had just searched for, 'may be associated with dangerous groups and individuals'. They were all too aware that any such initiative carries the risk of disturbing users and damaging trust. At the same time, causing friction in what might otherwise be a smooth user journey from white supremacist search to white supremacist group would be considered a success.

Ideally, this balance would be struck by ensuring that the safety module is only seen by genuinely high-risk users, and that therefore any disturbance or loss of trust is not felt by users who were in fact searching for something innocuous. While it is not possible to guarantee this through the use of keywords, clearly keyword selection was a critical component, both of the pilot programme's impact and for proof of concept.

Facebook provided Moonshot with the full list of keywords for each country which, during the pilot, would have triggered the safety module to appear in the user's results. To evaluate the keywords, we carried out three tests.

1. To test their relevance with white supremacist and/or neo-Nazi pages and communities on Facebook, we conducted manual searches on Facebook for all of them to see if they returned relevant results.
2. To test both relevance and breadth, we cross-checked Facebook's list with our own comprehensive database of violent far-right keywords, which include sections on neo-Nazism and white supremacy with USA- and Australia-specific keywords, alongside a wide range of other subcategories.[2]
3. Where keywords neither returned white supremacist or neo-Nazi results on Facebook, nor appeared in our in-house database, we conducted desk research to evaluate their relevance to white supremacy and/or neo-Nazism.

More than a year has passed since the creation of Facebook's keyword list in 2019 and our tests were carried out in mid-2020. During that time, Facebook removed a large number of pages from its platform.[3] Search results Facebook may also differ depending on the country in which the search is conducted in such a way that cannot be controlled for through the use of a VPN.[4] Our tests may therefore have returned different results than at the time of keyword selection in the respective location. Any violent extremist content found as a result of conducting these searches was immediately reported to Facebook.

UK-specific keywords and safety module error

The Exit Australia database primarily comprised keywords related to far-right extremism relevant to both the global and/or Australian context. Three keywords were specifically related to UK-based neo-Nazi and/or white supremacist groups and two were related to left-wing extremism. This was not a problem in and of itself - if high-risk individuals in Australia happen to be researching UK neo-Nazi organisations, it does no harm to present them with an alternative by way of an offer of support form local intervention providers.

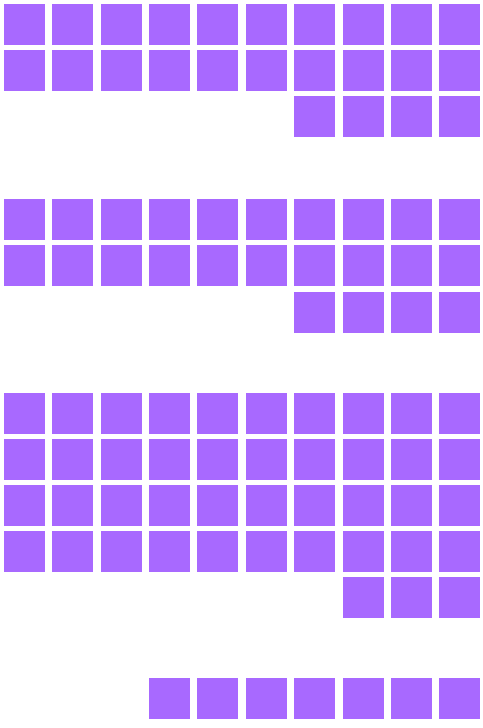
[2] Moonshot CVE's database of violent far-right keywords runs to approximately 30,000, each one coded by content type and risk. It has been tested and deployed for safeguarding purposes in, among other geographies, the United States, Canada, the United Kingdom, France, Germany, Australia and New Zealand.
[3] 'Facebook removes nearly 200 accounts tied to hate groups', ABC News, 6 June 2020
[4] Facebook uses a range of factors to determine user location and by extension, search results. Using a VPN to change user location, as one factor, does not necessarily override over factors, such as the self-identified location on a user's Facebook page.

However, we also discovered the safety module was appearing when these and other terms were searched for from a UK IP address. This was a problem, because this meant potentially high-risk UK Facebook users were being sent on an illogical user journey that ended with them being offered support from an Australian NGO. While we are confident that Exit Australia would have been able to provide help should anyone have made contact, nobody from the UK did refer themselves. This is not surprising, given the redirection will have undoubtedly have seemed accidental. It also helps to explain why 27% of the users redirected from Facebook to Exit Australia's website were from the UK, which will also have had an effect on the unusually high bounce rate mentioned later in the report (see page 14).

We informed Facebook of the unexpected safety module behaviour as soon as we became aware of it and the error was fixed during our evaluation.

Test results: keywords used to redirect users to Life After Hate

Of the keywords provided to us:



26%
returned white supremacist and/or neo-Nazi content on Facebook. Of the remaining 74% that did not,

26%
appeared in our in-house database of violent far-right terms;

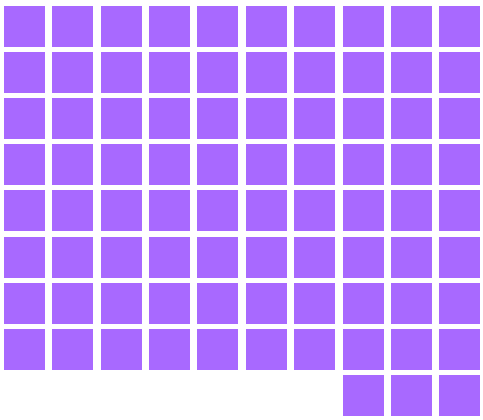
46%
were found to be relevant by desk-based research, and

2%
terms were of questionable relevance.[5]

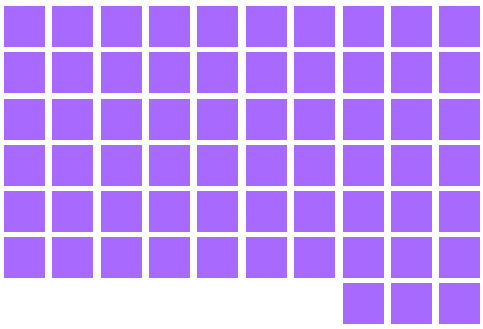
[5] Two terms were of questionable relevance to white supremacist and/or neo-Nazi movements: 'JewWIdder' and 'in blood and glory'. The first is probably a misspelling as it is not recognisable to Moonshot subject-matter experts or Life After Hate even as a word, irrespective of relevance. The second term is similar to 'blood and honour', the motto of the Hitler Youth. It is also the name of a (non-violent) video game, a film and, separately, a TV series about the Hitler Youth. Both should be double checked by Facebook for spelling and relevance.

Test results: keywords used to redirect users to Exit Australia

Of the keywords provided to us:



49%
were the same as those used in the Life After Hate list. Of the remaining 51%:



38%
returned white supremacist and/or neo-Nazi content on Facebook;

9%
appeared in our in-house database of violent far-right terms. Of the remaining 4% that neither returned relevant results on Facebook nor appeared in our in-house database:

3%
were found to be relevant by desk-based research, and

1%
keywords referred to far-left groups.

Stage 2: Did users who searched for these keywords click on the safety module?

Limits on data retention mean Facebook were not able to provide historic click-through rate (CTR) data for the pilot period but were able to provide contemporary CTR data. Between April and June 2020, the safety module averaged a click-through rate of 4%. This is high in comparison to CTRs for advertisements, including industry standards for digital ads (2%), advertisements used in Moonshot CVE's own Redirect Method programming (3-5%) and Facebook's anecdotal estimates of 4-5% for the highest quality advertisements. However, the safety module is not an advertisement, so it is difficult to say whether 4% is objectively high or low. We will return to this aspect of the evaluation in the final section.

Stage 3: Did users who clicked on the safety module interact with the delivery partner websites?

Over the course of the four-month pilot, Life After Hate and Exit Australia received a combined total of 2,288 visitors from Facebook. The rate at which these new visitors clicked through from Facebook to both delivery partner websites remained relatively consistent over the pilot period.

Compared to visitors from other sources, visitors from Facebook to both delivery partner websites were more likely to leave after viewing only one page, and spent less time on both sites overall.

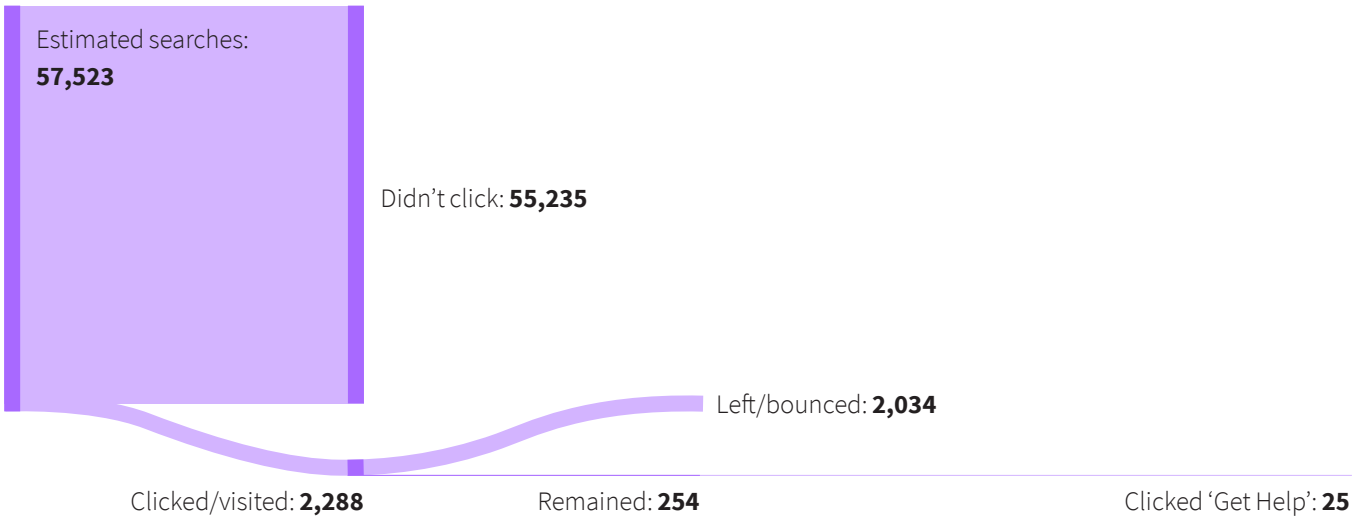
To some extent these results are to be expected. A user arriving on a website of their own volition is likely to spend more time on it than someone who was redirected to it without knowing what to expect. However, at between 85-89%, the bounce rates for each site are objectively high, both when compared to industry standards and, more usefully, when compared to other programmes run by Moonshot based on the Redirect Method in which users have ended up on a website for which they had not originally searched.[6]

As a counterpoint, there is value in these users accessing the delivery partner websites for any amount of time. Disengagement is a complex process which, in this context, may begin with the user simply being made aware of the existence of either delivery partner. That their first visit may not have resulted in their first outreach is not evidence of failure. The data show that the FRP is, at the very least, successfully introducing potentially vulnerable individuals to tailored support services, even if most are not immediately engaging.

[6] Bounce rates in commercial B2C user journeys average around 50%. Bounce rates in our own Redirect Method programming vary by context and platform, but average 60-70%. It is also important to bear in mind that comparison betweeen commercial user journeys and CVE user journeys is extremely limited, because they are attempting opposing outcomes; CVE user journeys are trying to disrupt and even invert the process of search query to expected/desired result.

Stage 4: How many users went from passive search to active conversation?

User journey flow from Facebook to delivery partner landing pages



On average, the number of users coming from Facebook who went on to click one of the delivery partners' 'Get Help' buttons was 5.75% lower than it was for users coming from all sources other than Facebook. As before, these lower percentages are to be expected given the difference in user intent; users not redirected to these sites via the FRP will most likely have made an active choice to visit the sites, many with the aim of accessing support.

In total, 25 individuals who initially sought to engage with violent extremism on Facebook ended up receiving some form of support from one of the delivery partners. This alone proves that, designed in this way, the programme can successfully link high-risk individuals with support, and that it can help to turn a passive search for violent extremist content into an active conversation.

The fact that a combined total of 25 individuals redirected by the programme chose to seek support from Exit Australia and Life After Hate means that, at this stage, we can conclude the following:

1. Facebook **can** use its search module to reach the 'low-prevalence, high-risk' audiences that advertising services are not designed to reach. In doing so, they have successfully:
2. created friction between a search query for a white supremacist and/or neo-Nazi community and a positive search result; and to some extent:
3. functioned as the conduit between high-risk individuals and their respective delivery partners, which has helped to turn some passive searches into active conversations.

Referring back to section II of the report, we can say that the three core components of the pilot have been shown to be functional and to some extent effective. The report now looks at the extent to which each of these variables could be made more effective, such that more of the right users see the safety module and click through to begin active conversations.

IV: Increasing the number of active conversations: observations and recommendations

The data from the pilot period shows that users redirected from Facebook to delivery partner websites:

- had in most cases never visited the websites before;
- were willing in some cases to begin active conversations with delivery partners, but
- spent less time on those sites versus other visitors, with 85-89 out of every 100 people leaving the site without browsing to another page (higher bounce rates, lower time-on-site);
- began fewer conversations with delivery partners versus other visitors (lower conversion rates).

In reality these findings will be the result of a combination of many variables, but fundamentally the three core components of the FRP - the keywords, safety module and landing pages - are themselves the three main variables that will have had the greatest impact on the results of the pilot.

During the pilot period, the users redirected from Facebook who did not go on to begin an active conversation with a delivery partner were, broadly speaking, either:

Curious about the safety module but not at-risk;
Relevant variable: keywords

At-risk but put off by the safety module;
Relevant variable: safety module

At-risk and not put off, but not incentivised to engage when landing on with delivery partner website.
Relevant variable: delivery partner landing page

This gives us three main variables around which to base our observations and recommendations.

Observations

I. There is an opportunity to expand the number of **keywords**

The pilot was conducted with 191 keywords in total, which we consider to be a sensible approach given neither Facebook nor the delivery partners could predict how many users would click through and request help. Furthermore, most of the keywords provided to us were relevant, according to our tests, and effective, according to the 25 users who began active conversations with delivery partners.

Nonetheless, 191 is a low number when considered alongside the vast corpus of neo-Nazi and white supremacist terminology. The Moonshot database of violent far-right words and phrases contains approximately 4,000 core terms similar in nature to the

ones used by Facebook.[7] There is, therefore, clear potential for expansion and we consider it likely that a measured expansion of the keyword database would lead to an increased number of Facebook users being redirected away from their search for white supremacist and/or neo-Nazi communities and toward active conversations with intervention providers.

Any expansion of keyword databases may result in increased delivery partner caseloads, so implications for resourcing need to be considered as part of any expansion initiative.

II. More data is needed about the impact of the **safety module**

The safety module can only be evaluated on its own merits as no comparable data from similar modules on either Facebook or other platforms was made available. Equally, while Moonshot has extensive experience redirecting vulnerable users away from violent extremist materials using advertising, the safety module is not an advertisement.

As previously mentioned, click-through rate data was unavailable for the pilot period of performance, but the module's current CTR is around 4%. This is relatively high compared to CTR data for adverts, but since the safety module is not an advert, such comparisons are of limited value. The differences between commercial adverts and on-platform modules might mean that, with testing, a 4% CTR could even be increased.

The safety module also bears some responsibility for the high bounce rates on the delivery partner landing pages. In a well-designed user journey of this kind, the landing page should fulfil the expectation created by the asset that directed the user to it, but we feel the FRP module text used in the pilot did not logically align with the overall user experience.

Here is the safety module copy again:

> *'These keywords may be associated with dangerous groups and individuals. Facebook works with organizations that help prevent the spread of hate and violent extremism. [Learn More]'*

Taken literally, 'Learn more' suggested users were going to learn more about Facebook's work with organisations that help prevent the spread of hate and violent extremism - which they were not. It also did not tell the user they were going to be taken off the platform - a major disruption to any user experience, even when a warning is given.

III. The FRP gives delivery partners an opportunity to tailor their **landing pages** and increase engagement

Few organisations have to cater to as wide a spectrum of website visitors as Life After Hate and Exit Australia. Visitors to a website that sells clothes might be male or female, young or old, but they all want new things to wear. Restaurant visitors might be vegan or carnivorous, but all of them are hungry. Trying to fulfil the needs of both a curious researcher and someone trying to escape a violent extremist movement within the same website is an unusual and extremely difficult challenge, but both organisations have had to try because like most sites, they don't know why a given user is visiting. The result is - because it has to be - generic homepages that cannot best serve the needs of the most vulnerable users, as evidenced in this case by high bounce rates.

[7] The full database runs to around 30,000 keywords but this number is achieved by combining core terms such as 'proud boys' with variations such as 'join' and 'donate.' All such variations are excluded for the purpose of this comparison.

Facebook Redirect changes this equation. Both organisations now know that all their visitors coming from the safety module are visiting because they searched for a violent extremist or hate-based term on Facebook's platform. This presents a significant opportunity to create dedicated landing pages or, ideally, microsites featuring content tailored specifically to these users' needs.

> *"I think our website needs to be restructured because it's very unclear who the target audience is - donors? Concerned family members? Formers? That needs to be front and centre. That is the number one thing that we currently need help with."*[8]

Recommendations

Design and process recommendations

#1 Expand the keyword list

To increase the number of Facebook users being redirected away from their search for violent extremist communities and instead, involved in active and necessary conversations with intervention providers, the keyword database should be expanded in cooperation with delivery partners - both for their localisation expertise and to allow them to resource appropriately.

#2 Test different safety module language and placement

As a result of this recommendation, Facebook has already adjusted the 'Learn More' call to action to make it clear that users will be taken off-platform. It was also discovered as part of ongoing conversations with Facebook that the safety module was not showing delivery partner logos alongside the CTA as originally intended. The addition of delivery partner logos alongside an improved CTA will likely have a positive effect on bounce rates.

The safety module is untested in the context of violent extremism, so testing should be conducted on the language used, tone of copy, and its placement. In particular we recommend testing alternative calls to action given the potential interpretations of 'Learn more'. The relative success or failure of any changes should be determined by increases or decreases in CTR, bounce rate and the number of conversations started on delivery partner websites.

We also recommend piloting the placement of the safety module in different areas of the platform, away from search. We recommend these as potentially valuable but additional experiments rather than replacements for the safety module as it is currently conceived.

a) The module could be placed in the 'similar groups' section of the sidebar of a violent extremist Facebook group. The alternative 'hard indicator' here would be that the user is browsing the page of a violent extremist group which they may have reached through search. They may be at greater risk as a result and this in turn may increase engagement with the module.

[8] From interviews with Life After Hate staff.

b) The module could appear as a notification sent to users who have searched for a relevant keyword, or engaged with a post containing a relevant keyword, similar to other initiatives deployed on the platform in response to COVID-19 disinformation.[9] Here, the hard indicator is the same (i.e. search), but the more personalised and 'private' delivery of the message might increase engagement. However, it would be important to test this assumption before rolling out any such initiative.

#3 Create delivery partner microsites and tailored content

Facebook should consider supporting delivery partners in the creation of separate microsites with content tailored specifically to the needs of individuals redirected from Facebook's redirect programme. By separating this audience from the delivery partners' other audiences (family members, press, job applicants), delivery partners would be able to focus the design of the microsites purely on having the most positive CVE impact by allowing users to complete their user journey that began on Facebook as smoothly as possible - for example, by immediately providing visible and clearly labelled 'Get help' assets and simple contact forms. Optionally, with very careful consideration and with Facebook's agreement, these microsites could also utilise Facebook Pixel to retarget users on Facebook.

Implementation recommendations

#1 Dedicate staff time at Facebook

> In the course of carrying out this evaluation, Facebook assigned a member of staff to oversee future onboarding and current and future delivery partner management.

All parties would all benefit from there being a dedicated Facebook staff member who is able to act as a Delivery Partner Manager, responsible for all aspects of the delivery partner relationship, including:

- the onboarding process, timelines and preparation periods;
- troubleshooting and other time-sensitive guidance during the launch phase;
- guidance on how to further improve the user journey based on Facebook data and delivery partner analytics, such as the terms receiving the highest number of searches.

This person would also need to act in a project management capacity as the internal bridge at Facebook between the otherwise disparate teams responsible for the FRP.

#2 Standardise the onboarding process

Create an onboarding pack for new delivery partners that contains:

- Facebook's top tips for best practice in UI/UX design, agnostic from subject or geography;
- Short case studies that cover the challenges and successes of other delivery partners;
- A set-up checklist (see Appendix I) to make sure the delivery partner is ready to receive referrals, including estimated numbers of new unique visitors their website is likely to receive;
- An introduction to their point-of-contact at Facebook and an overview of their role and responsibilities.

[9] 'Coronavirus: Facebook will start warning users who engaged with "harmful" misinformation', the Guardian, 16 April 2020

V: Conclusion and next steps

During the pilot period of performance, the Facebook Redirect Programme successfully:

1. Reached a portion of the 'low-prevalence, high-risk' audience;
2. Created friction between search queries for white supremacist and/or neo-Nazi communities and positive search results; and to some extent:
3. Functioned as the conduit between high-risk individuals and their respective delivery partners such that some passive searches became active conversations.

We have made criticisms, observations and recommendations throughout the evaluation process, several of which Facebook have already acted upon. Some of them were specific to the pilot and others are relevant to the FRP as it is currently delivered. Our Pilot Evaluation Report addressed the former in detail while this report focuses on the latter, as these findings remain most relevant.

This was a pilot of a new programme, with no full-time staff dedicated to it at Facebook and two delivery partners who were already providing intervention support services to large numbers of people. It is not surprising that there were problems of communication, timing and process. It is now critical that those problems and the risks they created are addressed and minimised, so that the programme can evolve from being functional to fully operational. Of those, the most important is the need for a standardised onboarding process. It is imperative that the FRP does not redirect high-risk Facebook users to delivery partners who are unprepared. It is hard enough to persuade anyone online to take the opposite user journey to the one they had in mind. For a high-risk individual to reach the point of starting a conversation and opening up to the possibility of change is hugely significant - it is vital they are not then let down by a foreseeable lack of resources. To this end, we have drawn up a sample Onboarding Checklist designed for use in future deployments (see Appendix I).

Since our evaluation began, Facebook has deployed its Redirect Programme in Indonesia and Germany, to reach users searching for violent extremist individuals and communities, and globally on both Facebook and Instagram to reach users searching for information related to QAnon.[10,11] We look forward to the continued evolution of the programme and to supporting Facebook in their efforts to connect high-risk users with information and services designed to protect both them and their communities from harm.

[10] 'Counterspeech: Redirect', Facebook Newsroom, undated
[11] 'An Update to How We Address Movements and Organizations Tied to Violence', Facebook Newsroom, 27 October 2020

Appendix I: Evaluating, onboarding and supporting delivery partners

Our core recommendation from this evaluation is that the process used to onboard future delivery partners needs to be standardised. This recommendation was informed by multiple findings as detailed in the report, but its fundamental purpose is to ensure delivery partners are ready and able to provide intervention support services to every user who may come their way.

We have synthesised our findings into the following checklist. While the specifics of future checklists may vary, they should be similarly designed to ensure that the provision of support promised to Facebook users by the programme is always deliverable in practice.

☑ **Timeline to launch agreed and finalised**

A go-live date has been agreed and communicated with enough lead time for the delivery partner to have completed this checklist, if not complete already.

☑ **List of keywords agreed and finalised**

Facebook has shared the final list that will be used to surface the safety module. Delivery partners are clear both about the list itself and the reasons for any changes to it.

☑ **Data provided by Facebook to delivery partners to help estimate caseload increase**

Facebook has provided delivery partners with qualitative case studies detailing the onboarding experiences of existing delivery partners. In particular this should cover increases in caseload and how they were dealt with, as well as lessons learned - for example, website alterations which, in hindsight, delivery partners would have made in advance of going live.

To complement the case studies, Facebook should provide delivery partners with an estimate of (a) search volumes for their pre-agreed keywords and (b) click-through rates, based on comparable historical data (we accept this will not always be available). Delivery partners should combine this data with their typical rates of conversion from website visitor to case referral in order to estimate any increases in caseload as a result of being connected to the FRP.

☑ **Delivery partner landing page is ready**

Delivery partners have created or optimised an existing landing page or microsite to receive visitors specifically redirected by the FRP, and its content meets the following criteria:

☑ **Google Analytics (or similar) is set up and recording data**
At a minimum, web analytics software is recording the number of new visitors, their bounce rate, user content flow and time on site. Ideally conversion tracking is enabled and connected to a case management system (see below).

☑ **Service offering is clear on the landing page**
The landing page or microsite clearly and immediately states how the organisation can help the user. There is a clear and obvious 'Get help' or equivalent button, contact form, phone number, chat function or other method of outreach.

☑ **Language is neutral and non-judgemental**
Delivery partners have checked their copy to ensure it uses neutral and non-judgemental language throughout. For example, "we can help" rather than "you need help".

☑ **Landing page contains supportive and engaging content**
The landing page or microsite also contains videos, articles and other media which deliver further information to redirected users who are perhaps not ready to make contact but could benefit from more information and support. This content should be relevant to the local context and the broad category of material for which the user was originally searching. For example, an FRP designed to counter white supremacy and neo-Nazism should be met with content relevant to those types of violent extremism.

☑ **A case management system is in place and sufficient training delivered**
Whether the delivery partner has in-house caseworkers or uses a referral approach, new referrals from the FRP are ready to be logged in a case management system with which delivery partner staff are comfortable and familiar. This will ensure all cases receive the support they need and, when combined with web analytics data, enable delivery partners to calculate the number of new referrals as a result of the FRP.

☑ **Final sign-off and launch**
Facebook staff have reviewed the delivery partner landing page or microsite, communicated to the delivery partner that the safety module will go live on the agreed launch date, and the delivery partner has confirmed receipt of this information.

Appendix II

Methodology

Moonshot developed and applied a bespoke evaluation methodology to the FRP. It was developed in a logic model format and was designed to assess three areas of the project: Project Process, Project Implementation, and Project Outcomes & Impact.

The evaluation assessed the project in reference to the stated objectives and aims of the initiative, and industry best practice.

To gather the necessary information and data, Moonshot conducted semi-structured interviews with staff at Facebook, Exit Australia and Life After Hate, all of whom were involved with the setup and deployment of the programme. Moonshot also received click-through rate data from Facebook and Google Analytics data from Exit Australia and Life After Hate.

Moonshot developed a number of research questions in order to frame and guide the evaluation approach, methodology and structure. These questions are grounded in the types of data and documents accessible for the purposes of evaluation.

Moonshot conducted in-depth, semi-structured interviews with staff from Facebook to assess:

- Their understanding and interpretation of what successful implementation of the pilot would look like.
- The extent to which project theory and objectives were appropriately developed.
- Whether appropriate project strategy, redirect design and timelines had been established.
- Whether communication between parties involved in the project was clear and effective.
- Whether there were any challenges or risks with identifying and onboarding delivery partners.
- Whether any adjustments, changes or improvements could be made to the project.

Moonshot conducted in-depth, semi-structured interviews with staff from Life After Hate and Exit Australia to assess:

- Whether appropriate project strategy, framing and timelines had been established.
- Whether communication between parties involved in the project was clear and effective.
- Their confidence and capacity with regard to the increase in case workload.
- The extent to which participants would benefit from training, and on which topics.

Moonshot was provided with keyword lists and click-through rate (CTR) data from Facebook as well as Google Analytics data from the websites of Life After Hate and Exit Australia. Moonshot then conducted quantitative and qualitative analyses in order to assess:

- The extent to which keywords returned results which were relevant and suitable in breadth.
- The suitability and efficacy of the call to action and the safety module, from a user perspective.
- The suitability and efficacy of the landing pages according to industry-standard criteria, such as immediacy of access to services; mobile optimisation; relevance of content; tone of copy and visuals; clarity of service offering.
- The measurable impact of the FRP.