ISD | Powering new generations against extremism

COUNTER CONVERSATIONS

A model for direct engagement with individuals showing signs of radicalisation online

Executive Summary

Executive Summary

Extremist groups deploy a clear strategy for radicalising and recruiting new supporters online: marketing their ideas through the spread of propaganda and then engaging interested individuals in direct, private messaging to recruit new members to their causes.

Direct engagement with radicalising individuals by mentors and 'intervention providers' is now a well-established component of offline counter-terrorism efforts in a number of countries. These programmes are delivered by both government and civil society, and often include former extremists and social workers as intervention providers.

Until now, online prevention efforts have largely focused either on the removal of terrorist content or on the production and dissemination of counter-narrative and counter-speech campaigns to compete with extremist propaganda. However, there have been no systematised attempts to supplement counter-speech efforts with direct online messaging and engagement at scale.

ISD's Counter Conversations programme is an experimental approach designed to fill this gap and test if the methods deployed in offline interventions can be brought into the social media domain. Delivered on Facebook to date and working across Extreme Right and Islamist ideologies, the programme provides an opportunity for individuals showing clear signs of radicalisation to meet and engage with someone that can support their exit from hate.

In this report, we present the findings of our most recent pilot programme of Counter Conversations. The results demonstrate the positive potential of direct online engagements and point to the need for further exploration into how this model can be deployed in a responsible, effective and scaled fashion, as part of a suite of online risk reduction methodologies.

Direct engagement with radicalised individuals

For the past three years, ISD has been working to appropriate the direct peer-to-peer messaging approach used by extremists and apply it to the online world.

This started with an initial small-scale pilot, conducted in 2015, using former extremists drawn from ISD's Against Violent Extremism (AVE) network to reach out online to over 150 Islamist and right wing extremists.[1] The results suggested that direct online outreach was a potentially viable tactic that was worthy of further exploration.

Accordingly, we developed a larger programme of action research designed to test the viability of delivering this work at scale in a secure and evidence-based fashion, and within a robust ethical framework.

With financial support from Facebook, this included:

- Developing a **semi-automated identification methodology** to efficiently yet carefully identify individuals publically supporting extremism and using violent language on social media platforms;
- Recruiting, training and testing the effectiveness of **different types of intervention providers**, including former extremists, survivors of extremist violence, and professional counsellors; and,
- Developing a r**obust risk assessment framework and safeguarding mechanisms** for undertaking direct outreach with individuals showing open signs of radicalisation online.

Identification

During the initial pilot programme in 2015, the identification of candidates for intervention was performed entirely on a manual basis by ISD experts and researchers. The first step in scaling direct engagement online was to develop a semi-automated identification methodology that carefully and accurately identifies individuals who are publically expressing signs of ideologically inspired hatred and violent sentiment towards others on social media.[2] This methodology consisted of:

1. Identification of Facebook accounts which were repeatedly engaging with public Facebook pages associated with the extreme right or Islamist extremism, or which tended to attract individuals expressing violently extreme viewpoints. In this fashion over 42,000 individuals were identified, an overwhelming majority of which were extreme right.

2. We then applied an approach that combined machine learning and a Natural Language Processing (NLP) algorithm to identify people who appeared to be using violent and dehumanising language against other groups of people on these pages.[3] This distilled the pool of individuals identified in step one to a group of around 7,000, of which a sample of 1,600 were analysed further.[4]

3. A process of manual review was then carried out by ISD experts and researchers on this sample of 1,600 individuals, gathering additional open-source information that indicated support for violent extremism including: profile pictures containing extremist imagery; likes/ positive comments on pieces of content supporting violently extreme groups; likes/ positive comments on extremist material; posts containing support for violently extreme groups; posts containing extremist material; friends within extremist networks; indication of offline involvement with extremist groups. This manual review also applied a risk-assessment framework based on well-established social work practices and purpose built for online interventions. Over 800 individuals were selected in this manner to be candidates for intervention.

Profile of Candidates for Engagement

Using open source data we were able to determine age, gender and other demographic details for individuals selected for online outreach. We found that:

- Islamist candidates were significantly younger than extreme right candidates: 3 out of 4 (72%) of extreme right candidates identified were over 45, while a slightly higher proportion (77%) of Islamist candidates identified were under 30.
- Female Islamist candidates presented the youngest age profile: 4 out of 5 (81%) were under 30.[5]
- Nearly 1 in 10 (9%) of male extreme right-wing candidates appeared to be current or former members of the armed forces

Intervention and Engagement

Just under three quarters of those candidates selected for online intervention (70%) were engaged by ISD's intervention providers, who initiated conversations through Facebook's application Messenger. The remainder were not engaged due to limitations in the number and capacity of intervention providers within this pilot programme, underscoring the need to train and professionalise intervention providers as well as explore technological innovation that can enable online outreach with greater ease and at greater scale.

Three metrics were used to consider the impact of online outreach: initial response rates, sustained engagements (conversations that included five or more messages between the candidate and intervention provider), and indications of potential positive impact during the course of the conversations.

Overall just under one in five (20%) candidates who were contacted responded, a significantly higher response rate than that usually seen with unsolicited email campaigns. Islamist candidates were more likely to respond,[6] with a response rate of one in four

(26%) compared with one in six (16%) for extreme right candidates.

Sustained engagement rates between intervention candidates and providers were achieved with nearly three out of four (71%) Islamist candidates and nearly two out of three (64%) extreme right candidates.

One in ten (10%) sustained conversations suggested the programme had a positive impact. This included candidates:

- Expressing an interest to take their conversation offline;
- Indicating that the conversations had challenged or changed their attitudes or beliefs;
- Suggesting that the conversations had a positive impact on their negative online behaviours.

We found that intervention providers were most successful in achieving sustained engagement when they responded immediately to a candidate who had replied to them, adopted a casual or meditative tone, and explicitly mentioned and discussed extremism.

Intervention Providers

Success rates of intervention providers varied. The most successful intervention provider achieved a response rate of 46%, with 83% of those interactions being sustained, and the longest conversation including more than 500 exchanges, demonstrating the best practice potential of providers.

The pilot was designed with a view to testing the viability of different types of intervention provider. The findings suggest that a variety of types of intervention provider can be successful:

- Professional counsellors were able to deliver more conversations than former extremists and survivors of extremist violence.
- Survivors of extremist violence were most likely to have a sustained engagement
- Former extremists delivered the fewest number of conversations over the course of the programme due to other professional responsibilities, but were the most likely to get an initial response.

Implications and Recommendations

ISD's Counter Conversations programme provides promising evidence that a solid proportion of individuals expressing support for violent extremism online can be identified at speed and scale and encouraged to engage with online intervention providers on a sustained basis.

There is potential to scale this form of online 'counter conversations' work across different social media platforms beyond Facebook, by expanding ISD's semi-automated identification technology and methodology, and scaling up intervention providers, including training professional counsellors so that they are confident to deliver this type of work, as well as increasing support for former extremists and survivors of extremist attacks through ISD's AVE Network.

There are however also significant risks inherent in online outreach work, particularly when targeting individuals already displaying signs of radicalisation. These must be taken into account in the design of any scaled, professional programme:

- De-confliction: A secure process of 'de-confliction' should be considered to avoid online outreach with individuals who are subjects of active police or security service investigations.
- Automation: While some form of automation is necessary to scale, it is vital that any automated identification process is supplemented with expert manual review in order to minimise the risk of outreach with 'false positives' (individuals accidentally or wrongly identified as being extremist).
- Operating in countries that lack human-rights-compliant referral mechanisms: Online intervention programmes should only operate in countries with human-rights-compliant referral mechanisms. If this is not observed then Intervention providers who operate in these countries may face legal risk, and candidates face exposure to potential human rights abuses including unlawful detention, torture or even extrajudicial killings by police or security services.
- Legal liability: Online intervention providers must be mindful of local laws that apply to having contact with known extremists as this could expose intervention providers to prosecution.

- Links with government-run programmes: The public is likely to view government involvement in online outreach as problematic Moreover, in many countries governments' presence and ability to undertake this work is severely constrained by legislation like the Investigatory Powers Act 2016 in the UK.[7]

With appropriate risk mitigation – including clear ethical frameworks and safeguarding procedures – it is possible to create an effective system of direct messaging interventions at scale. To do so we make the following recommendations:

- Leverage technology to achieve scale in both identification and intervention across multiple social media platforms:
- Further refine and develop semi-automated identification solutions across alternative social media platforms.
- Explore additional applications of technology to enable a triage-like system for initiating conversations, based on innovative programmes such as Crisis Text Line, and the use of video chat services for in-depth face-to-face engagement.
- Professionalise intervention providers through training, salary and pastoral support and use networks like ISD's AVE Network to provide a 'community of support' to formers and survivors interested in delivering interventions:
- Develop an accredited training programme and qualification for intervention providers.
- Pay intervention providers and provide pastoral support; ISD's AVE Network provides an ideal framework for this.
- Explore potential links with NGO-operated offline 'exit', disengagement and intervention programmes:
- Explore how ISD's identification methodology can be applied to facilitate referral to offline intervention programmes.
- Trial the effectiveness of 'counter conversations' across the radicalisation spectrum:
- Trial supplementing counter-speech efforts with direct messaging counter-conversations further 'upstream' (i.e. to audiences who are not yet showing strong signs of radicalisation but who may be at risk, for example by being within friend networks of individuals promoting an extremist ideology).
- Carry out further work to evidence and identify what constitutes behavioural or attitudinal change following intervention to further define 'success':
- Explore how technology can be implemented to track instances of medium-to long-term behavioural change following online intervention.
- Explore the considerations surrounding applying offline and online interventions in countries that do not have human-rights-compliant referral mechanisms for intervention:
- Explore, within international institutions and organisations, the legal and policy frameworks for delivering counter conversations in high risk areas in an ethical and legal fashion.

> "ISD's Counter Conversations programme provides promising evidence that a solid proportion of individuals expressing support for violent extremism online can be identified at speed and scale and encouraged to engage with online intervention providers on a sustained basis."

About this paper

This report outlines the results of a 12-month programme trialling a methodology for identifying individuals who are demonstrating signs of radicalisation on social media, and engaging these individuals in direct, personalised and private 'counter-conversations' for the purpose of de-radicalisation from extremist ideology and disengagement from extremist movements. This is the first programme globally which has trialled the delivery of online interventions in a systematised and scaled fashion.

To request a full copy of the report please email info@isdglobal.org

London | Washington DC | Beirut | Toronto

Designed by forster.co.uk. Typeset by Lookandfeelstudio.

About the authors

Jacob Davey is a Researcher and Project Coordinator at ISD overseeing the development and delivery of a range of online counter-extremism initiatives including the counter conversations programme. His research interests include the role of communications technologies in intercommunal conflict, the use of internet culture in information operations, and the extreme right globally. He regularly provides commentary on counter-extremism issues and has provided expert advice to national and local policy makers.

Jonathan Birdwell is the Head of Policy and Research at ISD. Jonathan oversees ISD's Strong Cities Network, a global network of mayors, policy makers and practitioners working to build community resilience to violent extremism, as well as ISD's Policy Planners Network. Jonathan also oversees all of ISD's research, including the setting up and running of ISD's Digital Research unit. Jonathan's research interests include the relationship between violent and non-violent extremist groups, cumulative extremism, digital literacy, as well as broader questions around political participation and trust in institutions. Prior to joining ISD, Jonathan worked for seven years at the London-based think tank Demos where he published over 40 research reports including The New Face of Digital Populism and The Edge of Violence.

Rebecca Skellett is a Senior Programme Manager at ISD overseeing online and offline interventions. Through ISD's Strong Cities Network Rebecca has been involved in the delivery of training to over 2,500 municipal level practitioners, building capacity in recognising and responding to signs of radicalisation. Previously, Rebecca worked on the front-line of the UK's Prevent CVE Programme across several London boroughs, and has extensive experience in overseeing individual case work, conducting community engagement, and developing local CVE programming, training and policy frameworks for local government and the education sector.