CHRISTCHURCH CALL

TO ELIMINATE TERRORIST & VIOLENT EXTREMIST CONTENT ONLINE

New Technologies and the Christchurch Call

Challenges, mitigations, and opportunities

This paper summarises key challenges new technologies are likely to present in countering terrorist and violent extremist content online, as identified by the Christchurch Call Working Group on New Technologies.

The paper also identifies key harm mitigation strategies, as well as opportunities these technologies present in countering TVEC. This is intended to guide Leaders in their discussion at the Leaders' Summit, and to present suggestions for solutions that could be prioritised by the Call Community over the next year.

A full issue report, prepared by the New Technology Working Group, is also available.

Generative AI

The rapid improvement of generative AI technologies has significant implications – positive and negative – for the work of the Christchurch Call.

Challenges

Generative AI, like all technologies, is at risk of exploitation by terrorist and violent extremist actors. For example, generative AI may be used to create large volumes of propaganda or misinformation for radicalisation purposes. It may be used to create fake instances of terrorist and violent extremist content, to recreate real-world attacks, or to alter existing content to support terrorist and violent extremist content. Terrorist and violent extremist actors may even create bespoke generative AI models designed to promote radicalising information.

Other risks posed by generative AI include the 'liar's dividend,' meaning a proliferation of fake content undermines societal trust in genuine content; and the potential risk of biased data leading Generative AI models to reinforce harmful – or even radicalising – rhetoric.

Harm Mitigations

A number of key actors have proposed methods of preventing or limiting the misuse of AI for nefarious purposes. Relevant ideas include:

- **Watermarking and labelling of AI-generated content**: including watermarks on AI-generated content may limit the radicalising impacts of artificial misinformation. It is, however, a partial solution – it positively identifies watermarked content, but does not provide assurance about other content, and TVEC actors are likely to find ways around watermarking – so additional tools will be needed.

- **Content provenance**: like watermarking, content provenance may be used to verify genuine content, thereby helping consumers maintain trust in its authenticity. This may help to limit the radicalising impacts of AI-generated misinformation.

- **Digital literacy**: improved public awareness about how misinformation and propaganda are used may help mitigate radicalisation risks by improving individuals' ability to critically evaluate information, including that created by generative AI.

- **Content detection:**

  This could include bringing together current datasets with on-the-stack AI solutions to innovate TVEC detection capabilities. While this approach builds on present-day capabilities, it also offers iterative learning opportunities that may be useful in understanding AI frontier model risks and opportunities. Concurrently, exploring AI foundation model-level TVEC detection possibilities may support global AI 'safety by design' discussions. Civil society organisations could partner with AI providers to help them understand the nuances of this work, ensuring it remains technically effective and rights-respecting.

- **Hashing of generative AI TVEC**: research into how TVEC can influence AI foundation models may also prove useful in informing future-state hashing initiatives that will need to cater to increasingly diverse content and data types. There are logical connections here between the Call's work on New Tech and the CCIAO's work, which offers opportunities to safely research and understand some of these new tools. AI providers, academics and civil society could collaborate to build an understanding of the current and future-state file types that will challenge existing hashing systems, with a view to developing a pathway forward.

Opportunities

Generative AI is also likely to offer opportunities more effectively to counter terrorism and violent extremism online. For example:

- **Improved detection of TVEC**: Generative AI technologies may be used to detect TVEC, with effects similar to existing hashing systems. This could help OSPs identify new TVEC in crisis situations, and/or look for existing or new TVEC that has been modified to evade hash matching databases. In line with OSPs' nine steps to tackle TVEC, announced alongside the first Call Summit in 2019, the Community could work to improve open-source access to new and existing AI-enabled tools aimed at detecting and removing TVEC.
- **Positive interventions**: Generative AI could potentially improve positive interventions by: better identifying effective intervention points; better identifying individuals who could be supported through positive interventions; converting positive interventions into other languages; developing new, personalised positive interventions; or scaling up interventions following a crisis event. There are also risks here – for example, if an intervention does not feel 'genuine' to the individual receiving it, it may further drive radicalisation. Governments, online service providers, partners, and civil society could collaborate to develop effective AI-enabled positive interventions.
- **Red teaming content moderation**: OSPs could use large volumes of AI generated content to red-team their platforms in order to identify flaws in content moderation systems, thereby improving their ability to remove TVEC while also preventing the inadvertent removal of non-harmful content.
- **Reducing the human cost:** Generative AI technologies may offer opportunities to automate aspects of content moderation tasks currently requiring extensive human input, reducing the exposure of moderators to TVEC and associated content.

Christchurch Call Initiative on Algorithmic Outcomes

- The Christchurch Call Initiative on Algorithmic Outcomes has initially focused on working with more traditional algorithmic applications – recommendation algorithms, machine learning – to enable safe research into the interactions between users and algorithmic processes.

- As the Initiative develops, it offers some promise in understanding the operation of generative AI, including ways to analyse the impacts of TVEC on large training datasets and on large models.
- The inclusion of key AI firms in the Call community offers the opportunity to build out this research, with the support of a multistakeholder community, and to improve the evidence base to support development of sound policy, technical, and regulatory interventions.

Immersive Tech

Challenges

Immersive technologies enable individuals to have increasingly realistic and interactive experiences online. Terrorists and violent extremists may exploit this to form social relationships with vulnerable individuals for recruitment purposes. They may also carry out virtual attacks in immersive spaces, or recreate real-world attacks in virtual environments. As the distinction between online content and online experiences blurs in immersive environments, content moderation is likely to become more complicated. These risks also apply in immersive gaming environments.

Mitigations

- **Evolving definitions of 'content':** Virtual recreations of terrorist attacks, or attacks on virtual targets, could be treated as TVEC for content moderation purposes. Similarly, online experiences like immersive social or gaming environments could be treated as hosting content. This approach would need to developed with the importance of protecting online freedoms at the forefront.
- **Expanding the Christchurch Call Community:** As immersive technologies become more common, we could prioritise onboarding OSPs that provide these services, including in online gaming contexts. This could improve community-based approaches, including enabling information flows to develop policy that can help prevent immersive TVEC from proliferating across multiple immersive platforms.

Opportunities

- **Positive community building**: While TVE actors may exploit immersive environments to form relationships with at-risk individuals, these environments also present an opportunity to build a positive sense of community online, thereby lessening the risk of radicalisation for isolated individuals.
- **Immersive positive interventions:** in addition to positive social relationships developing organically in immersive online environments, these environments may also offer new opportunities for targeted positive interventions. For example, former extremists could deliver or help develop offramps (e.g. to be delivered by a personable online avatar).
- **Gamified counter-narratives:** Immersive gaming may be used to further build positive relationships, or even to share counter-narratives in a more compelling way for at-risk individuals.

Web3.0

Challenges

Web3.0, the decentralised web, and blockchain technologies may be exploited by TVE actors seeking to store and share TVEC in areas of the internet where it is less likely to be subject to safety measures including content moderation. They may use decentralised platforms or blockchain storage as reservoirs to store TVEC that is removed from major platforms, and to then redistribute it via links on those platforms.

Mitigations

- **Moderating Search Indexes:** while decentralised platforms store content outside of a central host, search indexes for decentralised platforms are not. This offers the possibility that – depending on the nature of the content and the platforms – it is possible decentralised TVEC could be de-listed from search indexes, thereby limiting its reach.
- **Geo-blocking:** TVEC stored on decentralised platforms could also be geo-blocked in some countries or regions. This may raise be freedom of expression implications of this, depending on the content.
- **URL Hashing:** When content is stored on decentralised platforms, it may be linked to on major platforms frequently. To mitigate this, URLs linking to known TVEC in decentralised spaces could be included in hash databases, which could prevent the content being shared repeatedly on major platforms.

Opportunities

- **Blockchain-enabled content authentication:** content stored on the blockchain is extremely difficult to remove and is therefore useful for providing a reliable record of information. It may, therefore, also be useful in proving content is authentic, thereby helping to mitigate the 'liar's dividend' issue noted above.
- **Human rights protections:** While not an explicitly TVEC-related opportunity, it is important to keep in mind the significant human rights protections offered by decentralised web technologies. As it is difficult to remove content on decentralised platforms, users' freedom of speech can be well protected on the decentralised web.