CHRISTCHURCH CALL

TO ELIMINATE TERRORIST & VIOLENT EXTREMIST CONTENT ONLINE

New Technologies and the Christchurch Call

Key Challenges for the Christchurch Call Community

At the fourth Christchurch Call Leaders' Summit in September 2022, Leaders established a new workstream to consider how our multistakeholder Community can support the adoption of new technologies in a rights-affirming and safety-conscious way. They tasked work on:

- identifying how these technologies might be abused or misused by terrorist and violent extremist users;
- the challenges they present in countering terrorist and violent extremist content; and
- developing strategies to address those challenges.

This report responds to the first two parts of that tasking by identifying key technologies and the risks associated with them.

Introduction

At the fourth Christchurch Call Leaders' Summit in September 2022, Leaders established a new workstream to consider how our multistakeholder Community can support the adoption of new technologies in a rights-affirming and safety-conscious way. They tasked work on:

- identifying how these technologies might be abused or misused by terrorist and violent extremist users;
- the challenges they present in countering terrorist and violent extremist content; and
- developing strategies to address those challenges.

This report responds to the first two parts of that tasking by identifying key technologies and the risks associated with them. It has been developed in consultation with a multi-stakeholder working group to inform the development of recommendations centred around the opportunities these technologies present.

The purpose of this report is to guide the Christchurch Call Community as it develops policy recommendations to counter terrorist and violent extremist content online, including through positive interventions, technical solutions, and multi-stakeholder collaboration.

It is intended to help the Community and its Leaders understand the issues, and to inform our response.

The Christchurch Call to Action to Eliminate Terrorist and Violent Extremist Content Online plays a vital role in helping to make the internet a safer space. The Call addresses the use of the internet to exacerbate the fear, intimidation, and mobilisation that is both a cause and an effect of terrorist and violent extremist attacks.

The attack on March 15, 2019, demonstrated the extreme harms livestreaming can cause when used in the commission of a terrorist act - with millions of copies of the footage spreading virally across the internet. The world was unprepared for this kind of exploitation of what was, at the time, a relatively new technology.

Since 2019, the Christchurch Call has helped to improve the ability of online service providers and governments to prevent viral distribution of content like the Christchurch terrorist's livestream. The Christchurch experience highlighted the need to keep up with technological developments before terrorists and violent extremists exploit them.

It is important to acknowledge the important opportunities new technologies bring, and the need to avoid stifling innovation. The Community must consider the lessons learned from our experience in moderating terrorist and violent extremist content (TVEC) – without losing sight of successes to date. We should also consider the vulnerabilities of children and young people, who are often the earliest adopters of new technologies. Our work also needs to uphold fundamental freedoms, including freedom of expression, and contribute to a rights-affirming future for the internet.

This report is the Community's first step towards identifying opportunities and strategies to prevent terrorist and violent extremist exploitation of new technologies. It outlines some of the ways new and emerging technologies are likely to be exploited, and some of the opportunities for their positive use in countering terrorist and violent extremist content online.

The working group developing this report identified key risks associated with three areas of new technology: generative AI; immersive tech; and the decentralised web. This report is not intended to be exhaustive; the risks outlined are a snapshot of the challenges expected from these technologies.

Work on countering these challenges must recognise the societal benefits new and emerging technologies can bring. We have a valuable opportunity to use these technologies in a positive, safe way in countering the harms of terrorist and violent extremist content online. With the lessons learned through its experiences in moderating TVEC, and drawing on diverse perspectives, the Call Community can act to safeguard online futures against the harm of terrorist and violent extremist content online while upholding fundamental human rights and freedoms.

Generative AI and manipulated content

Introduction

Generative AI is already changing the ways we work, learn, and interact with information and with others. Generative AI is likely to have a wide range of positive uses – it may increase our productivity, reduce the costs of content creation, and improve access to education, information, and even health services. There are also exciting opportunities for the use of AI in preventing the proliferation of terrorist and violent extremist content online; the Call Community should proactively adopt these tools as we develop our next steps. Despite these likely benefits, AI-generated or -manipulated content will almost certainly be exploited by terrorist and violent extremist actors for radicalisation and recruitment.

Key Terms

Large Language Models (LLMs)	Large language models are foundational to AI technology like generative AI. They are trained on extremely large amounts of data, with billions of parameters, and utilise complex algorithms. They can understand, analyse, and even generate language and adapt to a wide range of tasks. Noteworthy risks of unintended LLM performance include emergence of unanticipated capabilities and so-called 'hallucination'.
Generative AI	Generative AI deploys algorithms on datasets for the purpose of producing new content – whether text-based or code (e.g ChatGPT), images (e.g. Dall.e or Midjourney), or other forms of content like videos, audio, or synthetic data. The content is generated in response to user prompts and feedback. Large models may exhibit emergent behaviour, such as the ability to solve programming problems, or unexpected gender, racial, or other biases.
Deepfakes	A deepfake is an image or video altered to put one person's likeness in place of another. Deepfakes are often in the form of videos, but deepfaked images and audio are also common. The use of artificial intelligence to create deepfakes helps to make them realistic and convincing.

Issues

Artificially generated propaganda

Terrorists and violent extremists may use generative AI technologies to produce large amounts of sophisticated content, including propaganda. The efficiency of generative AI models means this can be done quickly, at large scale, and with relatively little effort compared to manually created content. Propaganda can be made at scale for sharing by bots on social media, or to produce 'fake-news' about terrorist attacks. The ability to create large volumes of content through generative AI may also be exploited by adversarial actors seeking to test the boundaries of platforms' moderation policies and capabilities.

Existing commercially available generative AI models, such as ChatGPT and Bing AI, have safety measures and restrictions in place to prevent their misuse. It is likely – even inevitable –terrorists and violent extremists will create their own models, without the safeguards applied to publicly available models. It is also likely nefarious actors will train these models on intentionally biased or harmful information, leading them to produce harmful results.

Deepfakes and artificially generated audio-visual content

Deepfakes or artificially produced images and videos are also likely to be used for propaganda and radicalisation purposes. They may be used to create appealing content for propaganda. It is also likely terrorists and violent extremists will exploit the instinctive trust people put into what they see to make people believe prominent figures have said or done something they have not. This may contribute to increased social division and contribute to radicalising narratives.

As generative AI technology develops it is likely to become increasingly difficult to identify and debunk artificially generated images and videos. Examples of deepfakes and artificial images of individuals seen to date (such as the deepfake image of the Pope in a white coat, or the AI-generated hoax of a Pentagon explosion resulting in a momentary market slump) have featured figures of major global significance and have spread widely, enabling them to be quickly debunked. As these kinds of images become common, it is likely they will drive division at a local level where they are less likely to be debunked. Content known to be fake may also be harmful, as it can create emotional responses important to the radicalisation and mobilisation processes, regardless of whether audiences believe the content to be real.

The increasing prevalence of AI-generated images and videos may also lead to a deterioration of trust in genuine content. If prominent figures are photographed or filmed in a situation damaging to their reputation, they may claim the imagery is a deepfake, manipulated, or artificially generated, undermining the public's trust in genuine information. This 'liar's dividend' may lead figureheads for extremist movements to re-shape narratives to radicalise vulnerable populations or create a general lack of trust and sense of apathy across populations.

Bias

Not all radicalising or socially divisive AI-generated content is likely to be created by malicious actors. Generative AI models, whether they create text, images, or other forms of content, may contain harmful biases. For example, artificially generated content may contain stereotypes that contribute to radicalising biases. However, people may consider AI-generated content as objective – and therefore assume any stereotypes are based in truth rather than bias.

Opportunities

As the community develops strategies to address these harms, it will be important to focus on positive uses of generative AI. For example, there are educational benefits of generative AI. It may improve access to reliable information, thereby limiting the impacts of radicalising disinformation. There may be opportunities to use AI-enabled or generated positive interventions, AI-powered alternatives to content hashing, or to improve our ability to identify and remove TVEC using generative AI. Large volumes of artificially generated content may also be used to identify and repair gaps in existing content moderation processes.

Christchurch Call Initiative on Algorithmic Outcomes

The Christchurch Call Initiative on Algorithmic Outcomes, announced at the 2022 Leaders' Summit, is a project to develop tools to help the community fulfil its ambitions for algorithmic transparency. Working with non-profit OpenMined, the Governments of New Zealand and the United States, Twitter, Microsoft, and Dailymotion, the initiative is developing and testing tools intended to help the community engage researchers and analyse the interactions between users and algorithms, without prejudice to platforms' proprietary rights or user data privacy. Once tested in situ, the initiative's co-sponsors hope to work with the Call Community to deploy these tools to support greater transparency and understanding. This will also have benefits for understanding generative AI.

Immersive Online Technologies

Introduction

Immersive online environments enable increasingly realistic social interactions online. There are likely to be significant benefits for this – isolated individuals may be able to form valuable and genuine social connections online in a more effective way than is possible on current online systems. Terrorists and violent extremists looking to radicalise people online or promulgate TVEC are also aware of this capacity and likely to exploit it, potentially also using Generative AI tools as they do so.

Key Terms

Metaverse	The metaverse is a somewhat contested term. An inter-connected and inter-operable network of immersive online environments which will allow people to interact with one another much like how they do in the real world. These environments are expected to blend the physical and digital worlds and to be used for a range of purposes – from gaming to education to collaborative remote working. While 'proto-metaverse' style platforms already exist, these are currently 'walled gardens;' there is not yet a true, interoperable, hyper-realistic metaverse.
Augmented Reality	Augmented reality technologies combine digital content and the real world. This may be through a game on a smartphone that uses the phone's camera to overlay virtual elements onto the world around the user, or even through a headset that adds these virtual elements into the users' vision.
Virtual Reality	Virtual reality is like an extended version of augmented reality. Using a virtual reality headset, users are fully immersed in an interactive virtual world.

Issues

Recruitment and Radicalisation

The increased realism brought by social interaction in an immersive metaverse environment may make it easier for extremists to exploit vulnerable people seeking social connection and a sense of community, thereby increasing the potential to draw them into extremist spaces and the spiral of radicalisation. This threat is exacerbated by the use of avatars in the metaverse and immersive environments, which may make recruiters seem more familiar and trustworthy than in current online environments.

Recruiters may also be able to create virtual representations of real-world extremist leaders – dead or alive – and have these representations interact one-on-one with vulnerable people. This may help recruiters in making people feel connected to and inspired by violent actors, thereby contributing to their radicalisation. This may be quite convincing, especially when combined with artificial intelligence. The use of artificially generated conversation may even enable recruiters to radicalise people in this way at scale - without the need for human supervision of the avatars.

New forms of content

As the online environment becomes more interconnected and immersive, it is likely new forms of content will appear. The metaverse, for example, is likely to blur the lines between online experiences and online content. It is likely terrorists and violent extremists will exploit the development of new types of content and the associated blurring of lines, and TVEC in metaverse spaces will take many different forms. This is likely to exacerbate the already significant challenges online service providers face in moderating content.

The blurring of content and experiences online – and the blurring of online and offline worlds – creates a further challenge. As people begin to spend more of their time in virtual or augmented reality spaces, terrorists and violent extremists will likely seek to cause fear and psychological harm

by carrying out attacks on virtual targets with real significance to users. For example, an extremist could carry out an attack on a virtual place of worship, causing psychological impacts for those who worship there.

Terrorists and violent extremists may also exploit the hardware needed to engage in virtual worlds. They may, for example, be able to gain access to an individual's virtual reality headset. This could allow them to put illegal content in front of individuals against their will. It could also allow for physical harms such as damage to the ears or eyes through audio-visual attacks or even by causing seizures or otherwise creating adverse effects on users' brains.

Immersive Gaming

The potential risks posed by immersive online environments and TVEC may be seen through the example of immersive gaming.

There are existing examples of extremists using online gaming spaces to radicalise and recruit young people. This tactic has been used on Roblox, which is aimed at children and young people. There is also some evidence collaborative gaming is a powerful tool for extremist recruitment as it creates a sense of trust, community, and a common goal. Beyond recruitment and radicalisation, the community-building elements of collaborative gaming may also be used to increase feelings of solidarity among existing members of violent extremist communities. These risks are likely to be exacerbated by the increased realism of social interaction in immersive gaming environments.

As we have seen on existing sandbox-style platforms, it is likely the metaverse and immersive gaming environments will be used to gamify real-world terrorist and violent extremist attacks. We have seen gaming platforms like Minecraft and Roblox abused to recreate the Christchurch Terrorist attack, and harms of this type will likely only increase as online environments become more immersive and realistic. This gamification of real-world violence may be used to normalise, train for, and encourage extremist violence or to dehumanise victims of terrorist attacks.

Augmented reality technologies may also be used for real-world attacks – for example through the use of videogame-style 'score-counters' or targets overlaid on a video of a real-world terrorist attack. The Christchurch terrorist, and some of the attackers they inspired, livestreamed attacks from a first-person perspective, reminiscent of a first-person shooter videogame; the use of augmented reality to add gaming elements to such videos would likely contribute to the normalisation of extremist violence. This may also aid in the commission of attacks – supporters may be able to use AR elements of livestreams to guide perpetrators away from law enforcement or towards potential victims through, for example, the use of overlaid arrows on the perpetrator's AR goggles. In this way, the blurring between content and gaming also blends with the real-world commission of terrorist attacks.

There are other ways in which gamification of violence in this way may also help to enable real-world violence. Anders Breivik, the perpetrator of the 2011 attack on Utoya Island in Norway, claimed he used Call of Duty, a common first-person shooter videogame, to train for his attack. It is likely immersive gaming environments will increase the ability of violent extremists to train in this way.

Opportunities

While the increased ability to interact socially in immersive spaces is likely to be exploited for radicalisation, these digital environments may also allow socially isolated people to build positive social relationships – thereby making them less vulnerable to radicalisation. Immersive gaming and the sharing of narratives may also be used for good, by building communities around non-extreme views or by spreading playable counter-narratives.

Decentralized Web and the Blockchain

Introduction

The Decentralized Web offers great potential for the protection of human rights, privacy, and online freedoms. Users have greater control over their data on decentralised platforms, and it is more difficult for authoritarian governments to restrict speech on the decentralized web. These benefits also present opportunities for malicious actors to develop, store and spread TVEC with limited opportunity for the application of safety measures, including content moderation.

Key Terms

Decentralized Web (Dweb)	The decentralised web refers to the shift from online information being held by centralized servers and companies to being stored across a larger network enabled by peer-to-peer infrastructure. This potentially represents a shift away from a few large online service providers, like the major social networks of the 2010s, to a larger number of small online service providers.
Decentralized social network	The systems behind decentralized social networks are, unlike traditional social networks, made publicly available. This could allow individual instances of the networks to be created and moderated by individuals rather than a central owner.
Blockchain	The blockchain is a decentralized technology for storing information. Information on the blockchain is difficult to remove or modify. Blockchain technologies are often used for cryptocurrency transactions, but can also be used to govern platforms and software (see DAOs, below).
Peer-to-peer Connectivity	Peer to peer connectivity refers to the direct connection of individual devices online, without the need for a central server. While this is not a new phenomenon, P2P connectivity contributes to the growth of the decentralised web.
Decentralized Autonomous Organizations (DAOs)	DAOs are organizations whose rules are enforced by a decentralized computer program, often one stored on a blockchain. DAOs, like public corporations, may be managed by many partial owners, who vote or use other means to decide what the DAO should do. Unlike public corporations, DAOs may have unclear jurisdiction, as members and infrastructure may be broadly distributed throughout the world. DAOs may send and receive cryptocurrencies and software infrastructure, such as decentralized social networks or AI models.

Issues

Proliferation of platforms and peer-to-peer connectivity

When content is stored on decentralized servers – such as on the blockchain – it is extremely difficult to remove. This can protect the activities of groups that require privacy or need to protect the freedom to exercise their human rights online. It can also mean nefarious actors may use decentralized instances as “safety nets” or reservoirs for TVEC they expect would be removed from larger platforms. There is a high likelihood they will then these reservoirs to push content back into more mainstream online spaces. This is likely to enable increased proliferation of terrorist and violent extremist content.

This is exacerbated by the challenges with moderating or removing online content on the decentralised web. One of the key challenges is that product safety teams are less often found on decentralized platforms than on large platforms, and their work can be constrained by the operating environment. Those seeking to have illegal TVEC removed are therefore unlikely to have an

identifiable point of contact on these platforms, greatly hindering their ability to have this content removed.

There are also jurisdiction issues associated with attempts to regulate DWeb content. It is difficult to identify where content has been shared from and what laws apply – and therefore whether it is legal. This also creates difficulties with actioning takedown notices for content which is clearly illegal. It is also difficult to determine whether existing regulations in a given jurisdiction, such as the EU's Digital Services Act, with its focus on large platforms, will cover large networks of interconnected small platforms like the Fediverse.

TVEC is possibly also likely to be readily monetised on the decentralized web, where cryptocurrencies, which enable anonymised payments, are the primary form of currency. ISD notes Odysee, a blockchain-based social network which allows for monetisation of content via cryptocurrency, has been found to host copies of the Christchurch terrorist's livestream. The combination of difficulty removing this content and the opportunity to profit from its upload is likely to motivate individuals to upload TVEC on decentralised platforms.

Opportunities

The decentralised web and the blockchain also present some opportunities for preventing the spread of TVEC. For example, blockchain storage may be used to verify a piece of content, helping users to determine the authenticity of an image or video, thereby limiting the ability of radicalising disinformation to spread widely.

Next steps

Recognising the complex, diverse challenges emerging as this work is taken forward, the Christchurch Call's Working Group on New Technologies has agreed to work in three sub-groups – one for each of the technology areas identified above. These sub-groups are aiming to develop advice on countering some of the identified risks, including by building on the opportunities noted and identifying new ones. In line with the tasking from Leaders at the 2022 Summit, the Working Group will develop strategies for supporting the safe and rights-affirming adoption of these technologies, rather than stifling their development.

This approach will likely centre positive interventions and counter-narratives, recommending ways new technologies can amplify the effectiveness of these approaches, while minimising potential TVEC risks. It will be important to ensure recommendations are inclusive and consistent with the Call Community's commitment to upholding human rights and online freedoms.

For this reason, sub-group policy recommendations will be socialised across the broader New Technology Working Group, drawing on perspectives from the Call's multi-stakeholder Community. The Working Group drew on the opportunity of RightsCon to raise the issues presented by new technologies with interested participants, hearing from experts in human rights and technology. At this discussion, the clear message heard was that the Call's approach must be positive and forward looking, rather than restrictive.

We anticipate sharing initial policy recommendations, with a particular focus on identifying opportunities, at the Leaders' Summit in September 2023.

Thank you to the Working Group and Sub-groups for taking forward this important work. If you are interested in getting involved in one of these groups, please get in touch at info@christchurchcall.com

Annex: Key Christchurch Call Commitments

Commitment	Implications of new technologies
1. Counter the drivers of terrorism and violent extremism by strengthening the resilience and inclusiveness of our societies to enable them to resist terrorist and violent extremist ideologies, including through education, building media literacy to help counter distorted terrorist and violent extremist narratives, and the fight against inequality.	**Risks** Potential for generative AI and immersive tech harms to contribute to drivers of TVE including social division and deteriorating trust in information. Potential for generative AI and immersive tech to train and provide knowhow to TVE actors. **Opportunities** Uses of these technologies to create and share effective counter narratives, and for education/media literacy purposes. This may include technical solutions such as flagging or providing context for artificially generated content (e.g. Content provenance technologies, context notes on Twitter…). Working with young people as we do this to maximise efficacy for those most vulnerable
5. Consider appropriate action to prevent the use of online services to disseminate terrorist and violent extremist content, including through collaborative actions, such as: Awareness-raising and capacity-building activities aimed at smaller online service providers; Development of industry standards or voluntary frameworks; Regulatory or policy measures consistent with a free, open and secure internet	**Risks** TVE actors may use new technologies to create new forms of TVEC and spread these more widely and efficiently. **Opportunities** Work across Call Community to develop effective and rights-respecting ways of identifying and removing this content – for example through AI-enabled detection. Co-develop policy and regulatory recommendations for preventing terrorist and violent extremist exploitation of new technologies, while taking into account human rights and developing clear guidance and definitions of compliance, accountability, and risk mitigation measures.
6. Take transparent, specific measures seeking to prevent the upload of terrorist and violent extremist content and to prevent its dissemination on social media and similar content-sharing services, including its immediate and permanent removal, without prejudice to law enforcement and user appeals requirements, in a manner consistent with human rights and fundamental freedoms. Cooperative measures to achieve these outcomes may include technology development, the expansion and use of shared	**Risks** New forms of content and ways of distributing it may make it more difficult to stop the spread of TVEC. Risk of over-correcting and harming FOE. AI-enabled content moderation may lack transparency. **Opportunities** New AI detection models have the potential to be more efficient, accurate, and effective than things like hashing. We have an opportunity to feed into their development.

databases of hashes and URLs, and effective notice and takedown procedures.	
11. Review the operation of algorithms and other processes that may drive users towards and/or amplify terrorist and violent extremist content to better understand possible intervention points and to implement changes where this occurs. This may include using algorithms and other processes to redirect users from such content or the promotion of credible, positive alternatives or counter-narratives. This may include building appropriate mechanisms for reporting, designed in a multi-stakeholder process and without compromising trade secrets or the effectiveness of service providers' practices through unnecessary disclosure.	**Risks** Generative AI accelerates user journeys towards TVEC due to increased personalisation. **Opportunities:** The Christchurch Call Initiative on Algorithmic Outcomes is developing a platform for studying interactions between users and algorithms. This may also be useful in developing positive uses of AI. Opportunities to use recommender algorithms to connect users with positive interventions. Use of generative AI to develop interventions.
14. Develop effective interventions, based on trusted information sharing about the effects of algorithmic and other processes, to redirect users from terrorist and violent extremist content.	**Risks** Proliferation of content and networks, including via DWeb, may make it difficult to implement interventions. **Opportunities** The social uses of immersive technologies like the metaverse may have benefits. Individuals who are vulnerable to radicalisation due to social isolation may be able to find a safe sense of community in virtual spaces, thereby lowering the risk that they will become radicalised. Artificially generated positive interventions may be effective in driving users away from TVEC AI-enabled detection and provenance software may enable platforms to ensure that positive interventions and fact checking are reaching those consuming radicalising content.
15. Accelerate research into and development of technical solutions to prevent the upload of and to detect and immediately remove terrorist and violent extremist content online, and share these solutions through open channels, drawing on expertise from academia, researchers, and civil society	**Risks** Technology may develop faster than we can keep up with, making it difficult to ensure that technical solutions are effective. Or technical solutions may keep up – but harm online freedoms as they do so. **Opportunities** Utilise multi-stakeholder model to feed into the development of technical solutions as the technologies and challenges themselves arise. Build connections between trust and safety tech firms and Call Supporter platforms.