Christchurch Call Leaders’ Summit 2023: Compilation of supporting papers

This document contains papers produced in support of the Christchurch Call Leaders’ Summit 2023, including: the Joint Statement on the Summit’s outcomes; the Summit agenda and supporting papers for each agenda item; and the background papers reflecting the Call Community’s work over the course of 2022 and 2023 and in preparation for the 2023 Summit.
# Index of papers

**Outcomes of the Summit**

- 2023 Leaders’ Summit Joint Statement

**Meeting papers**

- Christchurch Call Leaders' Summit 2023 Agenda
- New Technologies and the Christchurch Call
- The CCIAO – Ethics and Governance
- Next steps for a resilient Christchurch Call
- Online Content from the Israel-Gaza Conflict

**Background papers**

- Concept Note: 2023 Leaders’ Summit
- Snapshot Report 2023
- 2023 Issues Brief
- New Technologies and the Christchurch Call
- Community Guidance: Government Transparency and Reporting
- Supporting smaller platforms
- ISD report: Misogynistic Pathways to Radicalisation
- Youth Engagement Session Concept Note
- Summit Co-Chair Statement 2022
Today, Leaders from Governments, Online Service Providers, Civil Society, and Partners of the Christchurch Call met in Paris at the Elysée Palace to continue our shared efforts towards the elimination of terrorist and violent extremist content online.

This year’s Summit takes place in the context of important challenges for the Call:

- The conflict in Gaza and Israel has given rise to significant volumes of terrorist and violent extremist content, and content that glorifies or promotes violence, antisemitism, and Islamophobia.
- Rapid advances in technology, including foundation AI models, present us with powerful new tools to better manage the problem of terrorist and violent extremist content online, and create new tools and vulnerabilities that can be exploited by terrorists and violent extremists.

Today, Leaders reaffirmed their commitment to a strong, resilient, and effective Call and to eliminating terrorist and violent extremist content, protecting and promoting human rights online, and a free, open, secure global internet as a force for good, and as a digital platform for innovation and social progress.

Our Community continues to expand. We welcome four new Call supporters: Anthropic, Discord, Open AI, and Vimeo; and seven new partner organisations: Extremism and Gaming Research Network, CASM Technology, MemeTech, Moonshot, Muflehun, Point de Contact, and Tremau. Our civil society Christchurch Call Advisory Network (CCAN) has also grown, with 11 new members. This expanded community brings new capability, expertise, and energy to the Call.

New Technology and its implications for terrorist and violent extremist content

Foundation AI models could transform the way we respond to terrorist and violent extremist content, providing better detection, prevention, and intervention capabilities, and the opportunity to scale and diversify safety tools in new contexts and languages.

Anthropic and Open AI join several existing Call supporters that have developed advanced AI capabilities. Important work is already underway to help realise this potential. This work includes a new collaboration between Microsoft and Tech Against Terrorism to pilot the use of advanced AI to enhance existing tools and, over time, make next-generation detection services available to smaller platforms.

Without safeguards, it is inevitable advanced AI capabilities will be weaponised by terrorists and violent extremists for propaganda, recruitment, and orchestration of attacks. Biases or vulnerabilities introduced through AI model training could present new vectors for harm or means of exploitation. Safety features may come with unintended adverse human rights impacts for diverse communities.

Red teaming and independent risk identification will form an important part of the toolkit to manage these risks. The international community has made progress in setting safety parameters and expectations around risk management, including through the leadership of individual Call supporters and the development of new forums and processes to help address overarching AI safety challenges.

Leaders endorsed the following actions:

- Working with the multistakeholder Call community to contribute to the development of frameworks, including in the context of the G7, GPAI, and AI Safety Summit, to identify, report, and mitigate terrorist and violent extremist exploitation of these tools;
- Engaging with the safety-tech sector and providers of commercial and open-source tools to support the availability of useful solutions to the sector, consistent with human rights and a free, open, secure internet;
Continuing to onboard relevant experts, service providers, and organisations into our community;

- Working on risk assessment and red-teaming efforts to mitigate the risks of terrorist and violent extremist misuse of AI drawing on the strengths and of our multistakeholder Call Community;
- Building on existing work by online service providers to align AI content governance tools with the Call commitments and facilitating more regular reporting and dialogue on these efforts;
- Drawing upon the expertise of the technical community to ensure our efforts support and do not compromise the maintenance of a free, open, secure, and interoperable global internet, and
- working with civil society ensure that human rights are promoted and upheld.

**Radicalisation and Algorithmic Outcomes**

The Call commitments seek to address the underlying drivers of terrorist and violent extremist content online, in addition to the more immediate issues of content moderation. They require that we consider the challenge of radicalisation to violence, which so often leads to the creation of new terrorist and violent extremist content.

Radicalisation is not solely an online phenomenon, but our work has highlighted the need for better insights about online activity, to enable the design of effective interventions, prevention measures, or other policy solutions. Users and communities play an important role in prevention and should be empowered to assist in this work.

The perspectives of young people are vital to understanding how the online user journeys of increasingly young perpetrators might contribute to radicalisation and exposure to terrorist and violent extremist content.

Last year Leaders asked the community to find ways of safely and effectively engaging with youth and reflecting their voices in our work. Today we hosted our first Youth Engagement meeting, where we heard directly from young people about their priorities in countering terrorist and violent extremist content online, and how they think we should go about this work. Leaders welcomed the initiative to launch a Youth Reference Group for the Christchurch Call to ensure young people can have input to the Call Community and its work year-round.

Leaders welcomed the report on *Misogynistic Pathways to Radicalisation* prepared by the Institute for Strategic Dialogue in association with the Christchurch Call and the Global Partnership for Action on Gender-Based Online Harassment and Abuse. Leaders noted the importance of taking action to address anti-LGBTQIA+ violence and gender-based violence as vectors of radicalisation, developing and implementing multistakeholder recommendations on this important issue and continuing to deepen the evidence base to support further action.

New regulatory measures to promote researcher access to data, and new voluntary transparency measures, including online service providers providing detailed information about the inputs to their content recommendations, and X’s decision to open source the inputs to, and the code that supports its content recommendations, have expanded the possibility to build public understanding and promote evidence-based decision making by policymakers, developers, and civil society.

Leaders welcomed the work of the EU Internet Forum including its 2023 Study on the role and effects of the use of algorithmic amplification to spread terrorist, violent extremist and borderline content, which provides an important snapshot of the different impacts of recommendation systems and user interactions and encouraged further efforts to address these risks.
The Christchurch Call Initiative on Algorithmic Outcomes (CCIAO), announced last year by New Zealand, the United States, Microsoft, and X, with the non-profit organisation OpenMined. CCIAO has made important progress towards building a privacy-enhancing system for sharing insights about algorithmic outcomes. In the first phase, the system has been built and tested, demonstrating it can safely and securely answer questions from accredited third-party researchers using information securely held by social media platforms.

Today, we announced France, Dailymotion, and Georgetown University’s Center for Security and Emerging Technology have joined the CCIAO, pledging additional funding of US$1.3m. This will further develop the CCIAO, to expand its functionality, and investigate and test its potential as a means of facilitating third party access, including in the context of new regulatory measures.

Leaders endorsed the following actions:

- The development of new governance structure and ethics frameworks for CCIAO, to enable its further development;
- Continued work to build CCIAO into a global network that has a significant impact for the Call, for ethical and responsible deployment of AI, and for the public good; and
- Specific efforts to improve the online experience of young people, and to counteract online misogyny and anti-LGBTQIA+ hatred as a vector for violent extremism.

Next Steps for the Christchurch Call

The Christchurch Call is a significant driver of positive change, operating as a multistakeholder community stewarded by the French and New Zealand Governments and underpinned by common commitments and values. Today, Leaders considered options for putting the Call on a resilient long-term footing that will position it for success in an evolving technological, diplomatic, and international regulatory landscape, and for drawing on the advantages of the Call’s innovative model to assist with related digital challenges.

Leaders endorsed the following actions:

- As co-Founders, New Zealand and France are committed to the long-term success of the Call. The co-Founders will work with the Call Community and potential funders to consider options and implement a structure for the Call secretariat that is resilient, better enables contributions from across our multistakeholder community, delivers results over the long term, and maintains the important attributes and values that have contributed to our success.
- The Call Secretariat will investigate partnerships with complementary initiatives that share the Call’s commitment to human rights and a free, open, and secure internet; to coordinate efforts towards eliminating terrorist and violent extremist content online; and to consider how work on related issues such as information integrity, the safe and responsible deployment of artificial intelligence, and youth radicalisation, can help enhance the fulfilment of the Call commitments, and to make it easier for small institutions, governments, and organisations to participate in multistakeholder efforts to shape technology norms.
- Looking ahead to the 5th anniversary of the Christchurch Call next year, Leaders will take decisions on the structures that are needed to support the Call as an impactful multistakeholder initiative into the future.

Crisis response and online content relating to the conflict in Gaza and Israel
Leaders acknowledged the fear and pain caused by the significant rise in terrorist and violent extremist content arising from the conflict, and related content inciting violence and hatred towards Jewish, Muslim, Middle Eastern, and all impacted communities and individuals around the world.

In the context of the conflict, governments and online service providers reaffirmed their Call commitments to eliminate terrorist and violent extremist content online, and to do so transparently and in a manner that protects and respects human rights and a free, open, and secure internet.

On crisis response capability, Leaders welcomed progress over the last year to bring more online service providers into the Call. They welcomed Tech Against Terrorism’s expansion of the Terrorist Content Analytics Platform, supported by funding from Canada, Australia’s announcement of funding for 24/7 crisis response capability which together improve support for small firms and close gaps in crisis response.

Leaders welcomed additional support to small firms, including through initiatives such as Altitude, a free, open-source tool developed by Google’s Jigsaw and Tech Against Terrorism, in collaboration with the Global Internet Forum to Counter Terrorism (GIFCT), to help platforms protect their communities from terrorist and violent extremist content.

Leaders endorsed the following further actions:

- Review our crisis response systems to ensure we can respond rapidly, effectively and in a rights-affirming manner to online crises involving the proliferation of terrorist or violent extremist content, including those arising from complex, protracted real-world incidents and conflicts, or involving bystander content.

- Update the Christchurch Call Shared Crisis Response Protocol, so that it can adapt to the new circumstances and challenges we face, harnessing the expertise and capabilities of partners and the Christchurch Call Advisory Network to monitor and address TVEC and human rights impacts.

- Consider and incorporate approaches for de-escalating tension and preventing on- and offline hate and violence, including strategic communications and positive interventions, in the updated Crisis Response Protocol.

- Build on work over the last year to deliver new tools and expand support for smaller firms to address TVEC on their platforms, including in crises. Integrate new capability with the existing infrastructure and protocols, including the Terrorist Content Analytics Platform (TCAP) and an updated Christchurch Call Crisis Response Protocol.

- Continue working with all sectors to test and refine the system, to strengthen human rights protections and increase transparency.

**Ongoing work in the Call Community**

The community has made important progress in a number of additional areas including; new guidance on government transparency; research efforts on the evolution of the threat landscape; greater coordination and linkages with initiatives such as the Freedom Online Coalition, Tech for Democracy, and similar initiatives to positively shape technology norms, and the application of aspects of the Call model by France in the development of its Child Online Protection Laboratory.

Leaders endorsed the following actions:

- In collaboration with the Action Coalition on meaningful Transparency, address barriers to transparency about work to eliminate terrorist and violent extremist content, including promoting the availability of information for independent research and assessment, and making use of our joint Transparency Initiatives Portal;
• Engage in dialogue on regulatory coherence to promote effective, complementary global approaches to the elimination of terrorist and violent extremist content;

• Reinforce the role of the technical community in the Call and to promote a free, open, and secure global internet;

• A multistakeholder process to benchmark and evaluate our work against the Call commitments and improve strategic oversight, transparency, and communication about the Call’s impact and value;

• Encourage the GIFCT to continue to align its efforts in cooperation with the Call community. Welcome its efforts to expand its membership, and encourage it to include a variety of platforms, such as smaller platforms and video games companies, to increase its impact in the tech industry;

• Continue to build meaningful engagement with civil society and other stakeholders, engage with and elevate diversity of views, cultures, regions, and experiences in our work, including with victim and survivor communities.

Conclusion

Four years on from the launch of the Christchurch Call, we have made considerable progress on our commitments and in preventing the weaponization of the internet by terrorist and violent extremist groups.

The digital environment now plays a central, critical role in the lives of people across the globe, and new digital tools deployed through the internet offer exciting new possibilities for advancement of social wellbeing and connectedness as societies.

The nature of terrorist and violent extremist content, and exploitation of online services is changing, and this will require us to adapt. We have a strong belief that the Call’s model, grounded in human rights, and a free, open, secure internet, and motivated by shared commitment to action, can continue to offer useful ways forward in a complex, contested digital environment.

Today we have shown that the Call continues to play a vital role in shaping the digital environment for good.
2023 Christchurch Call Leaders’ Summit Agenda
Friday 10 November 4-6pm

Notes on Summit Format: The Leaders’ Summit will operate under the Chatham House rule and aims to achieve constructive dialogue and interaction between Leaders from different sectors. Statements can be provided in writing for publication on the website but there will be no scripted remarks permitted in the meeting itself. Leaders will be seated around a table, with a limited number of accompanying staff watching from an adjacent listening room. President Macron and Special Envoy Ardern will issue a co-Chairs’ Joint Statement at the conclusion of the Summit.

1515-1530 Arrivals: officials and accompanying staff
1530-1550 Leaders arrive at venue
1600-1610 Opening Remarks: President of the French Republic H.E. Emmanuel Macron and Special Envoy for the Christchurch Call Rt Hon Dame Jacinda Ardern
A press pool will film President Macron’s and Special Envoy Ardern’s opening remarks, before leaving the room so the rest of the Summit continues behind closed doors.

Since 2019 the Call has established itself as a successful and ambitious multistakeholder initiative supported by a strong and growing community of governments, online service providers, partner, and civil society organisations. The Community is well placed to address evolving terrorist threats and online tactics, including in the context of real-world conflicts and as technological capabilities develop and change, to eliminate terrorist and violent extremist content while protecting and respecting human rights and a free, open and secure internet.

In reflecting on the Call’s role and mission, the President will highlight the responsibilities of platforms and service providers in relation to the issue of “active search”.

Special Envoy Ardern will invite a representative of New Zealand’s Muslim community to share a brief acknowledgement of the origins and purpose of the Christchurch Call, established following the March 15 terror attacks that took the lives of 51 worshippers in Christchurch, New Zealand.

The families of the victims of the Kouré attack, at the invitation and as special guests of the President, will briefly share their experiences and the importance of eliminating terrorist and violent extremist content online.

1610-1625 Part 1: Artificial Intelligence and the implications for terrorist and violent extremist content

The rapid development and deployment of advanced artificial intelligence models, present new challenges and opportunities for eliminating terrorist and violent extremist content online. Leaders will consider work done since the last Summit to understand these opportunities and risks. They will discuss how the Call must adapt from here. That will involve harnessing the Community’s technical capabilities and
expertise alongside the perspectives of impacted communities, to create effective new frameworks, tools, and interventions.

1625-1633  

**Part 2: Radicalisation and Algorithmic Outcomes**

Understanding and preventing radicalisation to violence, particularly for young users, is essential to fulfilling the Call Commitments. It involves online and offline factors. Leaders will consider the Community’s progress, including to understand user journeys and the impact of algorithms. The Christchurch Call Initiative on Algorithmic Outcomes (CCIAO) is breaking new ground, having developed and tested a safe, secure and effective research infrastructure. The Initiative is now moving to phase two of its work, including developing its governance and ethics framework. Leaders will identify next steps for preventing radicalisation, including scaling up use of the CCIAO infrastructure, to build the evidence base for effective policy-making and interventions in the online environment.

1633-1640  

**Part 3: Next Steps for the Call**

The Call is a unique, global multistakeholder initiative grounded in human rights and a commitment to a free, open, and secure internet. It has driven significant improvements in the ways governments and online service providers tackle terrorist and violent extremist content online. In 2022, Leaders signalled that interested Community members would consider how to support work on other, related issues. In this part of the discussion, Leaders will consider how to put support for the Call’s work on a resilient and sustainable footing, to enable supporters and the Community to continue delivering the commitments in the long-term. Leaders will also discuss how to apply the Call model to a small number of related issues with a material impact on delivery of the Call Commitments, such as information integrity and AI governance.

1640-16.55  

**Part 4: Israel-Gaza crisis and associated terrorist and violent extremist content**

The Hamas–Israel conflict has produced a sharp increase in the volume and complexity of terrorist and violent extremist content online, challenging trust and safety teams and measures, and the Call community’s incident response systems. It has spurred multiple incidents of Antisemitic and Islamophobic violence and is impacting communities around the world. Leaders will discuss how the Community is responding, and what more needs to be done to deliver Call commitments, deny terrorists and violent extremists a platform, and preserve online spaces for people to bear witness and for the media to report responsibly on events. Leaders will consider how crisis response should adapt, with emphasis on enhancing the Community’s situational awareness, de-escalating tensions online and offline, and preventing further violence and harm.

16.55-1700  

**Closing Remarks: Rt Hon Dame Jacinda Ardern**

As co-chairs, President Macron and Special Envoy Ardern will issue a Joint Statement at the conclusion of the Summit.
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.
The Christchurch Call Initiative on Algorithmic Outcomes: Ethics and Governance
The Christchurch Call Initiative on Algorithmic Outcomes: Ethics and Governance

Background

In September 2022 the New Zealand and United States Governments, Microsoft, and X launched the Christchurch Call Initiative on Algorithmic Outcomes (CCIAO).

Working with OpenMined, the CCIAO contributors proposed to develop and test a new system that would allow a trusted third party to answer critical questions about algorithmic outcomes, and trust and safety systems, using data from technology platforms, while using privacy-enhancing-technologies to ensure the underlying data and the privacy of users are protected.

This tool could enable accredited researchers in the Call Community to answer important questions about how recommendation systems impact users and what online factors play a role in radicalisation to violence. It could also help us assess and improve online positive interventions that help to reduce the risks of harm and violence. This could include study in specific areas such as gender-based online harm and its links to terrorist and violent extremist content and groups.

As it scales, the CCIAO tool could enable wider outcomes, including helping to unlock insights from new sources. It could help the Call Community study the risks and possibilities of foundation artificial intelligence (AI) models relating to terrorist and violent extremist content, facilitate regulatory compliance, and allow the multistakeholder Call Community to better understand the online environment in a holistic way.

Achievements in Phase One

In Phase One of the project, the CCIAO tool has been built, and put through multiple, rigorous proof-of-function tests. These tests have showed that the CCIAO tool can help mitigate the privacy, security, legal, intellectual property, and reputational challenges associated with enabling third party research on sensitive data. It is the first such system that has been tested and verified for use by multiple platforms and, importantly, is the only one we are aware of purposed for the public good.

The CCIAO tool has gone through an end-to-end multi-stage proof of function test at each of the three online service provider participants, using both synthetic and real data. It successfully completed this first phase of testing in August 2023.

This testing has proved the tool can operate in multiple environments and provide a consistent and secure process for answering defined questions from external parties. A small group of external researchers has been testing the system’s functionality and has managed to submit several successful requests. This important work will be presented at the workshop on Saturday, 11 November and positions us for the next phase of the project.

Next Phases and long-term vision

The tool is now ready to scale up – involving more organisations and bringing in more researchers. This requires development of detailed governance structures and an ethics framework for the tool.
This is important to achieve before scaling to ensure we are able to exclude potential malicious uses and other risks.

In Phase Two, the CCIAO proposes to:

- Establish a governance framework that supports multistakeholder outreach and fulfilment of the Call Commitments and principles;
- Build an agreed ethics framework that supports researcher access, enables better understanding of human/AI interactions, and research for public good;
- Investigate funded research partnerships with specialist experts, data custodians, and interested communities of practice;
- Investigate methods for assessing AI risk relating to terrorist and violent extremist content;
- Investigate a regulatory use case, as part of the European Union Digital Services Act;
- Continue to grow the number of contributors and the supporting infrastructure, with the aim of creating a global network of AI transparency tools for the public good.

The proposed long-term vision is for a broad network of CCIAO and similar technologies to be available across multiple institutions/entities, with a responsible governance framework that allows accredited researchers globally to analyse data in a privacy-protected way, across platforms. The network will be governed by an ethics and accredited user framework to enable safe, effective, access consistent with the shared objectives of the contributors and the Call Community.

**Elements of a Governance Structure and Ethics Framework**

A governance structure and ethics framework is important to ensure a means of governance and oversight of the initiative and future strategy; to enable CCIAO to grow within an agreed set of rules, ensuring core principles and ethics guidelines are upheld and the plan for Phase Two is effectively executed.

The Call will work with the Community to ensure governance and strategic oversight are appropriately managed and well-structured. This will include:

- Core principles to be applied to the CCIAO project, ensuring it is used for agreed public good purposes, that it protects the privacy of users and proprietary information held by platforms, and that it has an impact and helps to fulfil the Call commitments including consistency with human rights, transparency, and multistakeholder engagement;
- A governance board composed of the contributing organisations;
- Mechanisms to engage the Call Community and external experts on the fulfilment of the principles;
- Systems for bringing onboard new contributors, evaluating and adapting the network as it scales, and working with Call Community members to scope involvement;
• Exploring funding options for the next phases, ensuring well-documented, structured management of grants and contributions.

• Managing researcher accreditation and participation consistent with governance principles and the ethics framework, ensuring a diversity of expertise and experience.

• Identification and mitigation of risks to the CCIAO including e.g. from malicious misuse, unauthorised access, and other potential problems.

• Maintaining an adaptable, streamlined, scalable and effective project that meets the needs of the Call community and CCIAO contributors.

Next Steps
The CCIAO participants will hold a workshop for Call Community members in the margins of the 2023 Leaders’ Summit. This will provide an important opportunity for describe progress to date and show how the CCIAO works in practice. It will also provide an initial opportunity to describe the process from here on the governance structure and ethics framework, and the future steps for the project.

A more detailed paper will be circulated to the Call Community, drawing on initial feedback from the workshop, in the coming weeks.
Next steps for a resilient Christchurch Call

Supporting the Christchurch Call Community by building a resilient Secretariat and contributing the Call model to related digital challenges in the context of terrorist and violent extremist content.
Table of contents

Summary .............................................................................................................................................. 1
Proposal for a resilient Christchurch Call Secretariat ............................................................................. 2
Contributing the Christchurch Call model to the governance of related digital challenges in the context of TVEC ........................................................................................................ 4
Next steps ............................................................................................................................................. 5
Summary

Following Leaders’ direction at the fourth Leaders’ Summit in 2022, this paper identifies possible next steps for a resilient Christchurch Call.

This paper explores how the Christchurch Call Community can sustain and build progress towards its Commitments to eliminate terrorist and violent extremist content (TVEC) online with the support of a more independent, resilient Secretariat, while contributing the Call model to address digital issues as they relate to the elimination of TVEC online. Such issues potentially include those identified at the fourth Leaders’ Summit: information integrity and disinformation, new and emerging technologies such as generative artificial intelligence (AI), and young people and radicalisation.

Building greater resilience into the Christchurch Call Secretariat through establishing a structure with greater independence would grow the Secretariat’s capacity to support the Christchurch Call Community. The Secretariat is currently resourced by and embedded within the French and Aotearoa New Zealand Governments. Maintaining the current leadership by the Call co-Founders, while creating an independent legal identity for the Secretariat, could provide resilience through the ability to access input from a more diverse range of contributors, secure long-term funding, and deliver the scope to operate with greater pace, scale, and agility than at present.

The Call’s co-Founders seek Leaders’ support to build a more resilient and independent structure for the Secretariat. The Secretariat would be directed to work with the Call Community to implement this approach under the leadership and decision-making of the New Zealand and French Government Leaders, to put the Call on a sustainable path for the future.

This could expand the Secretariat’s capacity to meaningfully advance work as directed by Leaders at the fourth Summit, to apply the Call’s curated multistakeholder model to a small number of issues related to the elimination of TVEC online. Subject to Leaders’ direction, the Secretariat and Community, working with existing related initiatives and coalitions, could establish work programmes for some of these issues, including, as required, developing new stakeholder groupings, commitments, and projects that reflect the specific challenges these issues present.

Work on related issues and initiatives recognises the need to adapt to the growing complexities and changes in the Call’s environment, builds on the work of the Call to date, and enables fulfillment of Call Commitments, including to “[d]evelop and support a range of practical, non-duplicative initiatives to ensure that this pledge is delivered.”

The Secretariat has received positive indications of external philanthropic interest in supporting the Call on these proposed next steps.

---

1 Co-Chair Statement, Fourth Christchurch Call Leaders’ Summit, New York, 20 September 2022.
2 Call Commitment 24.
Proposal for a resilient Christchurch Call Secretariat

(I) Background

The Christchurch Call operates among a diffuse grouping of digital governance initiatives, and in an environment where digital challenges do not easily fit into existing governance mechanisms. The Call has become a significant driver of change, helping to build new trust and safety approaches and collaborations that are effective against the proliferation of TVEC online.

At the fourth Christchurch Call Leader’s Summit in September 2022, Leaders agreed the Call must remain focused on the scope of its Commitments, building on success to date, and sustaining this work into the future.

The Call operates in a rapidly changing, complex environment, where terrorist and violent extremist ideologies are intermingled with and enabled by other related issues, like the challenge of promoting information integrity and addressing the spread of disinformation, and technological developments, like increasingly sophisticated generative artificial intelligence tools. This environment has made it steadily more challenging for the Secretariat, within the limitations of its current structure and resources, to support the Call Community to address the Call Commitments.

To address the more difficult issues in the Call’s Commitments and work programme it is essential that Leaders and the Community have confidence that the support structure for the Community’s delivery of the Call Commitments is resilient to change and adequately resourced and sustained over time.

(II) The current Secretariat structure

At present, the work of the Christchurch Call is jointly led by the Governments of France and Aotearoa New Zealand, working with the technology sector, governments, and civil society leaders. This community of members is supported by a Secretariat function staffed by officials from the New Zealand and French governments.

Continued Secretariat support to facilitate progress on the Call currently relies on sustained political and financial support from these governments. The current structure limits the ease with which the Secretariat can access external funding (due to public finance requirements) and it is difficult to employ staff globally or from outside of the New Zealand and French governments, or to take on secondments or other inputs. This has implications for the scale and pace of the work the Secretariat can undertake itself or that it can support the Call Community to deliver.

(III) Requirements for a new Secretariat structure

The Call Community ideally needs an institutional support centre that is resilient to change, can be agile in the face of rapid developments and uptake of new technologies, serves as a model in global technology governance for smart multistakeholder solutions and ideas, and convenes the right people to deliver in a way that sustains and builds on the success of the Call, in accordance with Leaders’ direction last year.3

Call co-Founders, France and New Zealand, seek Leaders’ support at their 2023 Summit to develop options to place the Secretariat on a resilient, sustainable footing, as outlined in this paper.

With Leaders’ support and direction, the co-Founders propose to take subsequent decisions on options to restructure and redesign the Secretariat, and then direct the Secretariat to implement these

---

3 Co-Chair Statement, Fourth Christchurch Call Leaders’ Summit.
decisions, ahead of the next Leaders’ Summit. The Secretariat proposes to develop advice with options, working with the Call Community, for the co-Founders as they take these decisions.

Advice to the Call co-Founders on re-designing the institutional structure of the Secretariat will be guided by the overarching aim of delivering a resilient, sustainable Secretariat that can continue to, and improve, support to the Community and Call Leaders. Advice and decisions on a restructured and redesigned Secretariat will be guided by a set of critical criteria, including that the solution must:

- Improve delivery against Call goals by enabling greater pace, scale, and agility than at present.
- Be acceptable to the Call Community, including by enabling stakeholders to be represented in its design, and operation.
- Enable long-term funding from diverse sources; governmental and otherwise.
- Deliver ongoing New Zealand and French Government leadership as co-Founders of the Call, with the ability for strengthened leadership from other members across the Call Community.
- Maintain and strengthen interoperability with related initiatives.
- Remain sustainable in the long term, including being adaptable to changes in technology, the issues presented by TVEC, and political and commercial environments.

(IV) Possible Secretariat Structure

As an initial proposal, the Secretariat could be established, either wholly or partly, as a permanent, independent public-good legal entity (either a trust or non-profit organisation), based in New Zealand, France, or another suitable jurisdiction. A structure such as this, with careful design, could prove sufficiently flexible to meet the criteria outlined above.

The entity would seek funding for its operations and any agreed projects from a wide range of sources, including from Call Supporters and philanthropic funding. It would employ staff (or accommodate contributions of staff from Call supporters) globally.

A Board of Trustees or Directors would be established to oversee the Secretariat’s corporate and financial functions. The entity’s founding document would set the remit of the Board, and appointments of Board members (including maintaining the leadership of New Zealand and France as the two Call co-Founders).

The Board would not have authority over the Call work programme. Rather the work programme would continue to be set by Leaders. The Board would ensure that the Secretariat is performing well and that its operations are aligned with the policy directions Leaders determine at their annual Leaders’ Summit.

The Christchurch Call itself, as a curated multi-stakeholder community, currently has no corporate or legal structure and exists through the voluntary efforts, trust, and ongoing commitment of its members. The proposal for a new legal entity to provide the Call’s Secretariat function is not intended to fundamentally change the Secretariat’s current functions or the structure of the Call Community or its operating model. It is intended to significantly enhance the ability of the Secretariat to support and deliver on the Community’s intent. The restructured Secretariat would continue to support Call Leaders and the Call Community as it has to date, while also being capable of supporting any directions from Leaders to contribute the Call model to address digital challenges related to the Call’s mission to eliminate TVEC online.
Contributing the Christchurch Call model to the governance of related digital challenges in the context of TVEC

**(I) **Background

At the fourth Summit in 2022, Leaders discussed next steps for the Call and requests from some stakeholders to deploy the Call’s unique model in support of other digital issues as they relate to TVEC (‘related issues’). Leaders “agreed the Call should remain focused on the scope of its Commitments building on our success and sustaining this work into the future.” Leaders additionally agreed New Zealand, France, and other Call community members with an interest would consider some of these related issues, as they sought to understand:

- How we might apply what we have learned working on the Christchurch Call to related issues in the context of TVEC, such as disinformation and information integrity, harassment, abuse and hatred online, and issues affecting youth;
- How we can support and engage on related initiatives addressing these issues; and,
- Where multistakeholder interest lies in new work programmes separate to the Call.

In recognising the growth in new technologies, Leaders also directed the Community to launch a new stream of work to understand the adoption of new technologies in a rights-affirming and safety-conscious way to secure against TVEC. This timely workstream includes addressing the risks and opportunities of generative AI in proliferating and eliminating TVEC. This is one of the first instances of a multistakeholder community comprehensively analysing and finding solutions for generative AI challenges, which will be catalysed with the expansion of the Christchurch Call supporter base to include more generative AI companies.

These related issues have clear overlaps with the mission of the Call to eliminate TVEC online. The ability to address these issues in the context of TVEC will contribute to the fulfillment of Call Commitments to counter the drivers of TVEC by strengthening the resilience and inclusiveness of our societies, develop effective interventions, respect and protect human rights, and work across a range of initiatives to ensure the delivery of the Call Commitments.

The related issues are all significant digital and human rights challenges. The Call model could be meaningfully applied to drive smart solutions across these issues, either within the Call Secretariat structure or through effective partnerships.

Working on these issues as they relate to the elimination of TVEC online will advance the Call’s core work. It may also demonstrate that broader issues related to information integrity or generative AI can be best addressed through the multistakeholder approach and rights-affirming values demonstrated by the Call model.

---

4 Co-Chair Statement, Fourth Christchurch Call Leaders’ Summit.
5 Co-Chair Statement, Fourth Christchurch Call Leaders’ Summit.
6 Christchurch Call Commitment One.
7 Christchurch Call Commitment 14.
8 Christchurch Call Commitment 21.
9 Christchurch Call Commitment 24.
(II) Developing new work programmes

These related issues present specific challenges, and further work is needed to explore whether they require their own sets of commitments, projects, and stakeholder groupings to address the issues effectively while protecting human rights and a free, open, secure internet as a force for good. A strengthened Secretariat could work with the Call Community and support Leaders to take forward work to address these new digital, human rights challenges in relation to TVEC.

(III) Greater collaboration

The Call operates within a complex web of existing international initiatives and institutions. Existing global or institutional structures are generally poorly suited to emerging digital challenges, lacking the leadership buy-in, expertise, legitimacy, flexibility, or critical mass to effect change globally. At present, there is no natural ‘home’ for many technology governance issues. Multiple international institutions are working in this space without offering models that effectively promote the benefits of multistakeholder collaboration between government, industry, and civil society.

The Call model can play an important role in supporting Leaders to make meaningful progress on the most important digital challenges we collectively face. A strengthened Secretariat would also be better able to facilitate collaboration between the Call and other important initiatives working on issues related to TVEC through support and partnership.10

Drawing on "a range of practical, non-duplicative initiatives" in a coordinated and structured manner will better “ensure that [the Call] pledge is delivered.”11

(IV) Leveraging cross cutting capabilities

The Secretariat could continue to support the delivery of specific, funded technical outputs that can be applied across multiple digital challenges in the context of TVEC to deliver practical, credible solutions that inform changes in policy and practice. This includes, for example, leveraging the Christchurch Call Initiative on Algorithmic Outcomes to provide tools for enabling secure, privacy protected research on sensitive data, based on:

- a networked infrastructure with high levels of security;
- an agreed ethics framework that supports researcher access to enable better understanding of human/AI interactions, and research for public good;
- funded research partnerships with specialist experts, data custodians, and interested communities of practice;
- Multistakeholder engagement and governance.

Next steps

Leaders are invited to discuss the proposals in this paper for a resilient Call Secretariat, designed to support the Community’s delivery of Call commitments in a rapidly changing environment.

The Call’s co-Founders seek Leaders’ support to progress work to implement a resilient, independent structure for the Secretariat over 2023-2024. The Secretariat will work with the Call Community to

10 As in the Co-Chair Statement for the Fourth Christchurch Call Leaders’ Summit, such initiatives include Tech for Democracy, the Summit for Democracy, the Freedom Online Coalition, the Declaration for the Future of the Internet, the Aqaba Process, the Global Partnership for Action on Gender Based Online Harassment and Abuse, the Global Partnership on Artificial Intelligence, and the International Call to Stand up for Children’s Rights Online.

11 Call Commitment 24.
implement this approach under the leadership and decision-making of the New Zealand and French Governments.

Call Leaders are invited to discuss the role of the Call model in addressing digital issues related to the Call’s mission to eliminate TVEC and, should they wish to pursue this, to identify candidate issues and direct the Secretariat and Community to work on next steps to apply the Call model to these.
Annex A: An example model for a resilient Call Secretariat and work on addressing related digital issues in the context of TVEC

| Governance |
| Organisational Structure |
| TVEC issues |
| Commitment frameworks |
| Examples of potential delivery partners and collaborators |
| Shared technical solutions |
| Measuring impact |

**Christchurch Call to Eliminate TVEC - Leaders (providing Call direction)**

**Christchurch Call to Eliminate TVEC - Board (overseeing Secretariat functions)**

**Expanded Christchurch Call Secretariat (reporting to Call Leaders and Board)**

+ Multistakeholder Community

<table>
<thead>
<tr>
<th>&quot;Traditional&quot; Call issues, e.g., Content Mod. &amp; Crisis Response</th>
<th>Information integrity &amp; TVEC</th>
<th>AI &amp; TVEC</th>
<th>Youth radicalisation</th>
<th>Offline intersections with TVEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>The 25 Call Commitments</td>
<td>TBC</td>
<td>TBC, possibly related to US AI Voluntary Commitments</td>
<td>TBC</td>
<td>TBC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The existing Call Community</th>
<th>Declaration on Information Integrity</th>
<th>Frontier Model Forum</th>
<th>Call Youth Reference Group</th>
<th>Children Online Protection Lab</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOC Task Force on Trustworthy Information Online</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Christchurch Call Initiative on Algorithmic Outcomes**

Shared system for evaluation and reporting of progress
Online Content from the Israel-Gaza Conflict
Problem

The Hamas terror attack on Israel on Saturday, 7 October and subsequent Israeli military operations have led to the online dissemination of a massive amount of illegal terrorist and violent extremist content [TVEC], and harmful content that depicts and incites hatred and violence targeting Jewish and Muslim communities. While the ongoing situation makes analysis difficult, the volume and reach of illegal and harmful content appears unprecedented.

Online content generated during the conflict is having a profound impact on individuals, communities, and societies around the world. It is causing violence, fear, suffering, and damaging social cohesion.

Call Supporters and Partners have worked around the clock to respond. Based on open-source investigation and the research of Community members, however, it appears that the scale and complexity of illegal and harmful content has proved challenging even for the largest online service providers. This occurs in an environment where trust and safety teams and budgets have been pared back. The Community’s crisis response mechanisms have been challenged and gaps in these are apparent.

The situation raises important questions around how the Call responds to ongoing conflicts and converging crises that cause a spike in and sustained production of TVEC, including as part of distributed propaganda campaigns that raise the baseline of online hatred and violent content to dangerous levels.

The Call Community’s response must be sensitive to the views of impacted communities and responsive to their needs; protect and respect human rights – including to document and access information about significant events; and maintain a free, open, and secure internet.

Evidence

Call Partners and CCAN members have monitored, analysed, and reported on content emanating from the conflict. They have documented how videos and images captured by Hamas militants using head-mounted cameras and drones spread quickly from Telegram to all major social media platforms. Hamas and its supporters were also able to circumvent bans on platforms like Facebook by using victims’ own social media accounts.

Hamas content, material depicting Israel’s response and the ensuing conflict, and content of deniable or uncertain origin, has been picked up and repurposed in extremist networks, as part of conspiracies and in coordinated disinformation and foreign interference campaigns. This has contributed to a surge in antisemitic and Islamophobic hate and threats, and stochastic attacks against Jews and Muslims around the world.

---

1 GNET, Memetica, Tech Against Terrorism.
2 Brookings, Memetica.
3 Institute for Strategic Dialogue
4 Institute for Strategic Dialogue, Memetica, Tell MAMA UK.
5 For example in Beijing, Arras, Chicago and Brussels.
There is also a large amount of bystander and CCTV footage associated with the conflict and subsequent attacks. Often it has been shared by individuals bearing witness to these events and expressing their grief and outrage. Media have also incorporated both Hamas and bystander content alongside war footage in reporting, sometimes without warning labels or adhering to industry standards or ethics frameworks.

Response

Call Supporters, Tech Against Terrorism and the Global Internet Forum to Counter Terrorism (GIFCT) have worked hard to monitor and respond to the evolving situation, consistent with their respective policies and protocols. There has been a particular focus on preventing the threatened livestream or sharing of videos depicting hostage executions.

- For online service providers, the response has involved standing up special teams with Arabic and Hebrew language capability, making changes to recommendation algorithms and internal moderation queues, engaging third-party fact-checkers, and giving users advice and better tools to control their accounts and the content they see.

- The GIFCT activated the Incident level of its Incident Response Framework. The GIFCT has been working with members and stakeholders to monitor the evolving conflict and provide members with in-depth situational analysis, research, and bespoke knowledge products to support their teams.

- Tech Against Terrorism moved quickly to include the Hamas military wing in the Terrorist Content Analytics Platform, so that it could alert content to platforms hosting it. It provided intelligence reports and threat briefings to government and industry partners and issued public updates.

The sheer volume and variety of content, and different contexts and reasons for sharing have made it challenging to respond – from a legal, policy, technological and organisational perspective – and to do so in a swift, effective and rights-affirming way. Governments have called for more transparency around the industry response. Civil society organisations have expressed concerns about censorship, exclusion from access to the internet, and actions impeding the work of journalists and human rights defenders. They have also expressed concern about Government legal responses placing pressure on online service providers to over-remove content.

While the conflict continues there is little prospect that production and attempts to disseminate illegal TVEC and harmful content will tail off. Some commentators have suggested the situation raises the prospect of the online volume, reach, and intensity of TVEC and associated content settling at a new, much higher baseline than hitherto seen.

The Call Community must therefore act.

Options

It is critical that Supporters have in place the people, policies, systems and tools they need to deliver the Call commitments in the current

---

6 Tech Against Terrorism.
7 See, e.g., Meta’s and Google’s blogposts.
8 See GIFCT blogposts.
9 Access Now.
crisis, and to support smaller firms. Supporters must do so transparently, working in cooperation with Partners and CCAN.

This is an opportunity for the Call Community to look again at the crisis response system, to ensure it can be used to coordinate an effective response to the full range of incidents and content. That review should include consideration of situations where content from historic attacks threatens to spike, and the prospect of protracted conflicts generating complex mixes of content types. The European Commission’s work on bystander content is a useful contribution to this work.

The Christchurch Call Crisis Response Protocol occupies a unique place in the crisis response system. It can provide strategic oversight and direction to the system and a platform for transparency, dialogue and cooperation among Supporters, Partners, and CCAN. As part of the review of the crisis response system, the Community should update the CRP to strengthen these basic functions.

In line with Leader’s direction from last year’s Summit, this work should also consider how to harness the expertise and capabilities of the Call Community – especially among Partners and in CCAN – to monitor and share information about content and responses, (including adverse human rights impacts); to flag and report TVEC to Supporters, GIFCT and Tech Against Terrorism; and to provide open-source intelligence and analysis.

This is also an opportunity for the Call Community to explore how it can leverage its expertise and capability to better contribute to strategic communications initiatives and positive interventions aimed at de-escalating on- and offline conflicts where TVEC is a significant feature.

**Next Steps**

Leaders are invited to reaffirm the applicability of the Call commitments in the context of the terrorist and violent extremist content emanating from the Gaza – Israel conflict, and to discuss the proposals in this paper as means to strengthen the Call community’s response.

Subject to Leader’s direction, the Call Secretariat will work with the multistakeholder Call Community to review the crisis response system; to update and strengthen the Christchurch Call Crisis Response Protocol; and to identify where and how the Call can usefully deliver strategic communication and other positive interventions to de-escalate tensions and prevent harm. The updated Crisis Response Protocol and system could be tested in a multistakeholder tabletop exercise in conjunction with the five-year anniversary of the Call, with a report back to Leaders to follow by end of June 2024.
References

Access Now (2023) Civil society organizations’ call for tech companies to respect Palestinian digital rights in times of crisis and Precise interpretation of the DSA matters especially when people’s lives are at risk in Gaza and Israel


Global Internet Forum to Counter Terrorism (2023) How We Implement Our Mission and Our Incident Response Framework. Available online at: https://gifct.org/2023/11/01/how-we-implement-our-mission/

Global Internet Forum to Counter Terrorism (2023) GIFCT’s Ongoing Response to the Israel-Palestine Conflict. Available at: https://gifct.org/2023/10/13/ongoing-response/


Google (2023) Israel-Hamas war: $8 million in relief. Available online at: https://blog.google/inside-google/company-announcements/israel-hamas-war-relief/


Tech Against Terrorism (2023) Tech Against Terrorism alert tech platforms to online terrorist and violent extremist content from Israel-Palestine. Available online at: https://techagainstterrorism.org/news/tech-against-terrorism-alert-tech-platforms-to-online-terrorist-and-violent-extremist-content-from-israel-palestine


Tell MAMA United Kingdom (2023) 515 anti-Muslim cases reported to Tell MAMA between Oct 7-29. Available online at: https://tellmamauk.org/515-anti-muslim-cases-reported-to-tell-mama-between-oct-7-29/

Concept Note: 2023 Leaders’ Summit

In May 2019, following the 15 March Christchurch terror attack, Aotearoa New Zealand and France convened Heads of State and Government, Leaders from the Tech Industry and Civil Society to launch the Christchurch Call To Eliminate Terrorist And Violent Extremist Content Online.

Today, the Call Community consists of over 120 governments, online service providers, partners, and civil society organizations. We act together through the Call’s multistakeholder model to deliver on the Call’s 25 commitments, while protecting human rights and a free, open, and secure internet.

We have made real progress. The challenge of terrorist and violent extremist content is also changing fast, with the intermingling of violent ideologies and gender-based violence, growing concern about the radicalisation of young people, and rapid uptake of new technologies including powerful new AI models.

The Call offers a principles-based approach to dealing with these challenges, driven by Leaders from all sectors, harnessing the collective expertise, resources, and commitments of our community. In an age of contested visions for technology and society, the Call provides a way for us to face these challenges together and to make a positive impact.

Leaders from across the Call Community: Heads of State and Government, the Tech Sector, Civil Society, and Partner Organisations, will come together in in Paris at 1600-1800 on 10 November (CET) at the Elysée Palace, in association with the Paris Peace Forum, for our fifth Leaders’ Summit.

The Summit will provide an opportunity for a strategic conversation about our shared work to tackle terrorist and violent extremist content and deliver our vision of an internet for good. It takes place in the context of powerful new artificial intelligence models and a rapidly changing regulatory and geopolitical landscape.

This meeting will be an unscripted, forward-looking multistakeholder engagement among Leaders on an agreed set of topics. As with previous Summits, we will focus on ensuring ample input from each sector. This will be a chance to announce significant outcomes, talk about our successes and some of the pervasive problems we face, and chart the path forward for our community.
Summit Details

Leaders from the Call Community will meet in Paris on Friday 10 November 1600 – 1800.

- For Governments, attendance will be restricted to invited Heads of State or Government.
- For Online Service Providers and partner organisations, attendance will be restricted to invited senior executives.
- For Civil Society, the secretariat will work with the Christchurch Call Advisory Network to ensure a senior level of participation and diversity of representation and expertise.

The meeting will operate under the Chatham House Rule and will be independently moderated.

There will be no provision for pre-recorded statements or prepared remarks in the meeting. We do encourage statements submitted in advance as part of the ‘success stories’ process (see below) or by email to plenary@christchurchcall.com for inclusion in our communications.

Space is limited, and attendance in person will be restricted. The Summit will be livestreamed to the wider Community. Registration details will be circulated to the Community shortly.

Preparing for the Summit

We have worked with the whole community to shape the Summit including: areas where we have made progress; specific deliverables, individual or collective successes; issues that require more work or further discussion, and things that require strategic direction from Leaders.

In our ‘save the date’ message we signalled the following key themes as the starting point for this effort:

- The implications of new technology, particularly foundational AI models;
- Assisting small firms with eliminating terrorist and violent extremist content online and responding to crisis incidents;
- The evidence base on gender-based hatred and radicalisation to violence;
- Youth, algorithmic outcomes, and pathways to radicalisation;
- The Christchurch Call Initiative on Algorithmic Outcomes;
- Sustaining and expanding the Call Community.
- Multistakeholder approaches to related issues.

To ensure everyone had a chance to contribute and help shape this event, we designed and ran a process to gather input from our multistakeholder community across multiple channels, including:

- Inviting each Supporter and Partner, and members of the Advisory Network, to answer our survey on where the Community has made progress over the last year and where there is more work to do. Many of you are doing important Call-related work, including individually or in other fora. The Summit is operating under the Chatham House Rule, so the survey is a key opportunity for you to highlight specific contributions, successes, and
priorities. Your statements will be collated into a Snapshot Report for Leaders and published on the Call website. The survey will be open for completion until 16 October (NZDT). You can find the survey [here](#) and an example of how each response will be presented [here](#).

- As in 2022, we worked with you to produce an ‘Issues Brief’ that reflects our shared view of the Call’s progress, identifies areas where we need more focus our attention, and suggests possible ways forward. The Issues Brief formed the basis for identifying agenda topics for Leaders, and for New Zealand and France to develop the co-Chairs Statement.

- ‘Town Hall’ Community meetings in late August and early September to discuss progress, key themes, and issues that require Leaders’ attention.

- We also held Community Engagement Sessions to onboard new Online Service Providers and Partner Organisations under our workstream to sustain and expand the Call Community. We will hold further Community Engagement Sessions throughout September and October for prospective Online Service Providers and Partners.

### Issues for Leaders

Drawing on input from the Community, we have identified three issues for discussion by Leaders at the Summit itself:

- Jointly consider the impact on our shared mission of powerful new and emerging artificial intelligence models, the opportunities and risks they represent, and the actions that we will take as a result.

- Consider efforts to improve the user environment, address radicalisation and harm, and enhance researcher access to data, including assessing progress on the Christchurch Call Initiative on Algorithmic Outcomes, which is opening the way to better research and public understanding of AI.

- Support and sustain the multistakeholder model of the Call, grounded in human rights and a free, open, and secure internet, into the future in a complex global environment. This may include how we apply a similar model to related issues such as information integrity, AI safety and ethics, protection of children (with possible inspiration from initiatives such as the Children Online Protection Lab) and young people, and the future of the internet.

### Opportunities for further input from the Community

<table>
<thead>
<tr>
<th>2023 Survey</th>
<th>Four simple questions:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Where has the Call made progress over the last 12 months?</td>
</tr>
<tr>
<td></td>
<td>What have been your particular contributions and successes?</td>
</tr>
</tbody>
</table>

Open Here ([links to mailchimp survey form](#))

Please complete by 16 October

Example response [here](#)
Where have you faced challenges?
Where does the Call Community need to focus in the year ahead?

To be published as an attributed statement.

| Town Hall Meeting | These will be an opportunity to discuss and refine Summit outcomes, based on the final Issues Brief. | Sessions tailored to your time zone the week of 16 October |

**What to expect**

These inputs will be incorporated into the Summit agenda, and reflected in communications that take place on 10 November.

<table>
<thead>
<tr>
<th>Finalised Issues Brief</th>
<th>to inform participants and stakeholders</th>
<th>Sent to Community 9 October</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Snapshot Report</strong></td>
<td>Containing your survey responses and highlighting success stories and other points of interest and importance.</td>
<td>Sent to Community ahead of the Summit, and published on Summit Day (10 November)</td>
</tr>
<tr>
<td><strong>Co-Chairs Statement</strong></td>
<td>Issued by NZ &amp; France, drawing on the issues brief</td>
<td>Published 10 November (Summit Day)</td>
</tr>
<tr>
<td><strong>Community Meeting</strong></td>
<td>We will organise a meeting with the Community following the Summit to discuss the outcomes and how they will be implemented over the coming year</td>
<td>TBC week of Summit</td>
</tr>
</tbody>
</table>

**Other Events and Meetings at the Paris Peace Forum**

The Paris Peace Forum gathers a global multistakeholder community to discuss and find solutions to major challenges of our time. One of the four key themes for this year’s Paris Peace Forum is ‘Ensuring trust and safety in the digital world’. Specifically, the Forum will bring together key players in generative AI to outline a system of standards that would respond to the main risks posed by this new technology. You can find more information about the Paris Peace Forum [here](#).

In addition to the Forum’s existing programme of events and meetings, we plan to hold several side meetings during the Forum to support the policy outcomes of the Leaders’ Summit. These will include a Town Hall meeting following the Summit, and sessions to cover key topics such as the Christchurch Call Initiative on Algorithmic Outcomes. We will provide more information shortly.

**More questions?**

Please contact plenary@christchurchcall.com
This report compiles responses to the Christchurch Call Community Survey
# Table of contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>0</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>1</td>
</tr>
<tr>
<td>Germany</td>
<td>3</td>
</tr>
<tr>
<td>Japan</td>
<td>3</td>
</tr>
<tr>
<td>New Zealand</td>
<td>4</td>
</tr>
<tr>
<td>Google</td>
<td>6</td>
</tr>
<tr>
<td>Microsoft</td>
<td>7</td>
</tr>
<tr>
<td>Global Community Engagement and Resilience Fund (GCERF)</td>
<td>8</td>
</tr>
<tr>
<td>European Commission</td>
<td>8</td>
</tr>
<tr>
<td>Tech Against Terrorism</td>
<td>9</td>
</tr>
<tr>
<td>Cyber Threats Research Centre (<em>CCAN member</em>)</td>
<td>11</td>
</tr>
<tr>
<td>Federation of Islamic Associations of New Zealand (<em>CCAN member</em>)</td>
<td>12</td>
</tr>
<tr>
<td>Global Internet Forum to Counter Terrorism (<em>CCAN member</em>)</td>
<td>13</td>
</tr>
<tr>
<td>Institute for Strategic Dialogue (<em>CCAN member</em>)</td>
<td>16</td>
</tr>
<tr>
<td>InternetNZ (<em>CCAN member</em>)</td>
<td>17</td>
</tr>
<tr>
<td>Internet Society (<em>CCAN member</em>)</td>
<td>18</td>
</tr>
<tr>
<td>United Nations Counter-Terrorism Committee Executive (UN CTED)</td>
<td>19</td>
</tr>
<tr>
<td>Global Project Against Hate and Extremism (<em>CCAN member</em>)</td>
<td>20</td>
</tr>
</tbody>
</table>
Introduction

Ahead of the fifth Leaders’ Summit on 10 November 2023, the Secretariat invited Christchurch Call supporters, partners and members of the Christchurch Call Advisory Network (CCAN) to outline their contributions to the Call over the last year, and to share their views on where the Call Community as a whole has made progress and where it needs to apply more effort.

This Snapshot Report presents responses from a cross-section of the Call Community. It includes governments and tech companies, industry bodies like the Global Internet Forum to Counter Terrorism and Tech Against Terrorism, research institutions, civil society organisations and community advocacy groups.

The responses from the Call Community underscore the strength of the shared commitment to the Call. They also show the many and varied ways that Call Community members are taking the Call’s work forward, individually and collaboratively.

The responses describe the range of work under way deliver the Call commitments. They include the work of CCAN members to provide research and advice to governments that are looking to develop and implement regulation – like the EU Regulation on Terrorist Content Online, and the Digital Services Act; the creation of practical tools to help tech companies, including smaller ones, identify and moderate content in line with their own policies and new laws; ongoing efforts to monitor, report and action violative content, including in crisis situations; and dialogue and guidance on presenting issues like borderline content.

The responses recognise solid progress against the priorities identified by Leaders at the Summit in September 2022. There has been a concerted effort to raise the profile of the Call internationally, build the Call Community and CCAN, and create opportunities for dialogue. Through targeted engagement in focused working groups and projects, the Call Community has made tangible progress, producing a range of useful outputs. They include reports on the risks and opportunities presented by new technologies, addressing gaps in support for smaller firms, improving transparency around governments’ role in content moderation, and steps to understand and address gender-based violence online.

The Snapshot Report contains a clear message that the Call Community must continue working on longstanding priorities. This includes crisis response, extending and adapting our protocols and systems to close gaps and meet new challenges. Another longstanding priority is to understand the impact of algorithms on user journeys. The Christchurch Call Initiative on Algorithmic Outcomes has seen Call Community members build and test infrastructure for remote research on platform data without compromising security or privacy. This initiative offers the greatest promise for real insights into the role of terrorist and violent extremist content (TVEC) in radicalisation – alongside borderline content, conspiracies and disinformation – particularly for youth. Respondents consistently identify an urgent need for further work in this area to understand the problem and to find and apply effective interventions.

Respondents highlight the uniqueness and value of the Call’s genuinely multistakeholder approach, and the effort made to include the voices of impacted communities. The work needs to continue, with more opportunities for cross-sectoral dialogue around how governments and tech companies are delivering their commitments and how they can improve. This is one way to achieve meaningful transparency and accountability, which will also require the Call Community to find new and genuinely multistakeholder ways of evaluating impact, building on CCAN’s pilot project in 2022-23.
Czech Republic

This statement was provided by the Ministry of the Interior of the Czech Republic, the Ministry of Foreign Affairs of the Czech Republic, and the Office of the Government of the Czech Republic

Overall, where has the Christchurch Call made progress in the last 12 months?
We are glad to see the appeal on being future-proof and reflective of new developments in the online sphere. We are also very glad to see the emphasis on transparency as well as work with the younger generation. The latter is especially important to truly understand how to protect some of the most vulnerable ones online, while understanding their online behaviour.

This is also very pertinent for the online gaming environment. At the same time, the younger generations can bring new ideas at the table and should become an integral part of our discussions forward. It is equally important to keep engaging smaller platforms. They do not face the same obligations under some of the current pieces of EU regulatory legislation as the large or extra-large online companies and can therefore become a target of nefarious actors as well as it may be harder to make them compliant in cases where compliance is not in their goal.

We commend the continuous efforts to make the activities of the Christchurch Call community interoperable and to not duplicate activities. We also appreciate the emphasis on gender-based violence and extremism, since last year has once again reminded us that this area needs to remain high on our agendas.

Last but not least, we appreciate the bilateral outreach and a visit by New Zealand representatives in Prague and the opportunity to discuss the Christchurch Call activities and possible future needs.

What have been your particular contributions/successes?
The Czech Republic held its six-month Presidency of the Council of the EU in the second half of 2022. Online threats have been high on the CZ PRES agenda. Worth underscoring are two Presidency events that both focused on threats online: Online Dimension of Counter Terrorism and Nazi Memorabilia and their misuse for anti-Semitic ends. The former also hosted a presentation on the Christchurch Call to Action. Under the CZ PRES, the Council approved general approach on the proposal for Regulation laying down harmonised rules on artificial intelligence. EU member states thus took an important step towards the approval of ground-breaking and pioneering rules for the use of artificial intelligence. The final version of the text is now being negotiated with the European Parliament.

The Czech Republic has also continued to cooperate in relevant fora, such as the EU Internet Forum. There, it actively worked towards including borderline content, disinformation, and conspiracy narratives into the scope of the Forum’s activities. It also supported the creation of the EUIF Handbook of borderline content and contributed towards its content. Online gaming is also our long-term priority, which was reflected also during the CZ PRES 2022 event on terrorism online in Prague.

Domestically, the Czech Republic prepared and passed an implementation law thanks to which the Regulation on addressing the dissemination of terrorist content online is implemented in the Czech law. Domestic non-legislative activities include an active cooperation with online platforms and with the gaming industry.

In 2022, the foundation of Strategic Communication (StratCom) teams was established across four significant ministries, prominently within the Office of the Government. The efficacious implementation of StratCom teams across governmental institutions has laid the groundwork for a robust and enduring communication channel between the government and its citizens.
Where have you faced challenges?
Overall, the online spaces have been a fertile ground for conspiracy narratives and disinformation that also contribute to radicalization. There are online communities formed around these phenomena and some of their members also meet in real life. The tendency has been to use less popular and therefore also less regulated platforms.

Radicalization online of “lone wolves” is considered an ongoing threat. The past months, similarly as across the EU, the Czech Republic has seen a rise in extremist activity, including also the activity of so-called anti-system actors. We have also seen cooperation of “traditional” radical actors with this anti-system scene that does not quite fit into the right / left wing frame. Some of these anti-system actors have managed to translate their online activities successfully into opportunistic offline mobilisation at manifestations.

It is important to note that the Czech Republic has been a target of Russian disinformation campaigns for years. We currently observe a surge in community radicalization, notably orchestrated by anti-system and pro-Kremlin entities, resulting primarily from the deliberate online dissemination of misinformation and propaganda. Recent violent confrontations, some with fatal outcomes, between the emergent Ukrainian war refugee community and the historically marginalized Roma minority underscore the imminent challenges we anticipate in the ensuing year. The escalation of animosity, fuelled by disinformation agents within communities, holds potential to expand from virtual platforms to tangible confrontations. Overall, the Russian war against Ukraine has had a significant influence on the amount and scope of these activities and related effects on society and selected criminal activity and has had a significant impact on the online sphere, including condoning this aggression. It is important to highlight that the courts deal with these activities.

Where does the Call Community need to focus effort in the year ahead?
The Czech Republic sees clear benefits in the cooperation set up thanks to the Christchurch Call. We would like to see more work done on how the borderline content, disinformation, and conspiracy theories contribute to radicalisation that can lead to violent extremism, or terrorism.

We would also like to underscore the importance of work with the gaming industry. This industry may still be less well understood than other parts of the online sector, yet the number of gamers and games available makes it clear there is a need for effective cooperation and deep understanding. That also relates to the urgent need to keep engaging young users of online services to help the policy makers in better understanding of the newest trends and behaviour of the users.

Of ongoing concern are alt tech platforms and unmoderated platforms. We are optimistic that the full application of DSA will bring positive changes, however, the need to remain vigilant.

Last year has also marked a rapid growth in AI and its availability to wider public. New technologies, immersive technologies, and artificial intelligence need to be well understood. As for AI, we must work on ensuring its appropriate use by the relevant actors, but we should also understand it as a possible tool in the hands of criminals.

In general, it is important not to stay behind and devise approaches that are future-proof, technology and ideology agnostic, and while respecting human rights and freedom of speech, are flexible enough to help tackle online threats. Our last appeal would be to concentrate more on other language realities than just English. There is an urgent need of the technological companies to have language skills and local know how to be able to indeed effectively enforce their terms of service and contribute to the fight against terrorism and violent extremism. It is no longer possible to underestimate or otherwise not focus of the English-speaking world.
Germany

This statement was provided by the German Federal Ministry of the Interior and Community

Overall, where has the Christchurch Call made progress in the last 12 months?
We welcome the work of the Christchurch Call over the past year. The great advantage of the Christchurch Call is that it brings together a wide range of stakeholders with different perspectives. We especially appreciate the progress made on different topics in the working groups.

Germany is particularly interested in the dialogue within the working group on new technologies. For example, the last meeting “Christchurch Call Community Discussion: On AI-Enhanced Web Search and the new Bing” provided great impetus for future discussions.

What have been your particular contributions/successes?
In the EU, the TCO Regulation entered into force on 7 June 2021 and became applicable on 7 June 2022. The main and innovative feature of the Regulation is that the competent national authorities may issue cross-border orders requiring internet providers offering their services in the EU, including those based in another EU member state, to immediately remove terrorist content. Providers must comply with these orders within one hour of receipt. Germany has already transposed the Regulation into national law and cooperates closely with the EU Internet Referral Unit (IRU) and the competent bodies in the EU member states.

The Digital Services Act (DSA) entered into force on 16 November 2022 and will become applicable throughout the EU on 17 February 2024. However, very large online platforms (VLOPs) and search engines must start to apply the DSA at the end of August 2023. The DSA sets universal horizontal rules on due diligence obligations for intermediary services such as online platforms, among other things. It also includes measures to counter illegal content – a term that is not defined in the Regulation. Reference is made in this regard to Union law and legislation of the member states.

Where have you faced challenges?
We continue to see a large amount of jihadist propaganda. The sheer volume shows that we need to continue to work on reducing such content. This responsibility lies mainly with the companies.

Lastly, possibilities for anonymisation continue to play a major role with regard to violent extremist and terrorist actors online. For this reason, platforms that do not cooperate with governments are still very popular among many of those actors, who use the platforms to create networks and spread propaganda. Due to the lack of cooperation with the German authorities, it is extremely difficult or impossible to remove such content and prosecute the creators.

Where does the Call Community need to focus effort in the year ahead?
In our view, the added value of the Christchurch Call is the opportunity for dialogue with a wide range of stakeholders. In particular, this dialogue should continue to be used to improve cooperation between businesses and governments with regard to removing illegal content and reporting subscriber information to identify offenders. In Germany’s view, we can only fight illegal content effectively if the content is removed and criminals are prosecuted.

Japan

This statement was provided by the Government of Japan

Overall, where has the Christchurch Call made progress in the last 12 months?
We extend our appreciation to the Christchurch Call for its remarkable commitment in countering violent extremism online. While providing a pathway for the responsible and safe use of cutting-edge
technology, the Christchurch Call has provided an open forum for discussion where countries, companies and civil society can freely share insights, knowledge and strategies for countering violent extremism online.

We note in particular the outcome of the Transparency Working Group, which has provided a solid insight into understanding the online content ecosystem from a transparency perspective, where there had been less discussion of government transparency in relation to content moderation.

The New Technologies report should also be highlighted as an important first step for the Christchurch Call to identify opportunities and strategies to prevent terrorist and violent extremist exploitation of new technologies.

What have been your particular contributions/successes?
In Japan, following the adoption of the Christchurch Call to Action, we established a public-private study group on online exploitation. The Ministry of Foreign Affairs, the Ministry of Internal Affairs and Communications, the National Police Agency, private companies and NGOs regularly discuss how technology companies should take action against terrorism and violent extremism.

This year, in partnership with GIFCT, we will convene a workshop, inviting representatives from the wider sector to strengthen engagement and partnership, and to identify potential opportunities for collaboration. In addition, this year, Japan joined the Christchurch Call Transparency Working Group to further contribute to the community. We shared our approach and commitment to government transparency in content moderation.

Where does the Call Community need to focus effort in the year ahead?
While acknowledging Christchurch Call’s substantial achievements, we believe there is an opportunity to broaden its impact even further by expanding its membership, especially in the regions of Africa and South East Asia. In these regions, companies face unique challenges and their inclusion of Christchurch Call’s membership would enhance the global effort to counter the terrorism and violent extremism online.

New Zealand

This statement was provided by the New Zealand Government

Overall, where has the Christchurch Call made progress in the last 12 months?

The Christchurch Call has continued to provide leadership in support of the commitments to eliminate terrorist and violent extremist content (TVEC) online while continuing to advocate for human rights and a free, open, secure internet.

Over the past 12 months the Call has:

- As part of the Christchurch Call Initiative on Algorithmic Outcomes (CCIAO) and together with the United States, X, Microsoft, OpenMined, DailyMotion and LinkedIn, the Call has built and tested a system for remote research. Proof of concept has been successful achieved, including rigorous testing of privacy tooling and consideration of the legal, ethical and policy aspects of the project. Georgetown University’s Center for Security and Emerging Technology (CSET) has provided further support with a specific focus on on-chip transparency. The project team are now looking at a new stage of CCIAO development with the Call Community, which is building appropriate governance and ethics frameworks.
The Call Community launched a New Technology Working Group – to build its understanding of the risks and opportunities at the nexus of new technologies and TVEC. There is a strong consensus from across the Call community towards devoting time to consider how, for example, foundation AI models, may affect the effort to eliminate TVEC online. Key risks that have already been identified as part of this research include generative AI, immersive technologies, and the decentralised web. The working group is now developing advice and strategies to mitigate these risks.

The Christchurch Call Secretariat is in the process of setting up a Youth Reference Group to provide advice to the Call community. A youth engagement plan is being developed to ensure the safety of participants and enable them to make a meaningful contribution.

Working closely with Tech Against Terrorism and experts in CCAN, the Call has developed guidance to improve transparency around government supporters’ laws, policies and actions against terrorist and violent extremist content online. The Call has also supported the Action Coalition on Meaningful Transparency to develop its Transparency Initiatives Portal, to connect members of the community to existing resources, reports, and upcoming events, and support multistakeholder efforts to improve both technology company and government transparency.

What have been your particular contributions/successes?

New Zealand takes its role as co-leader of the Call very seriously, as it does its own commitments as a government supporter.

In March 2023, Prime Minister Hipkins appointed Rt Hon Dame Jacinda Ardern, former Prime Minister of New Zealand and co-leader of the Call, as the Prime Minister’s Special Envoy for the Christchurch Call. This appointment reflects Dame Jacinda’s personal commitment to the work of the Call and the ongoing strong backing of the New Zealand government. Working closely with France, Dame Jacinda’s relationships with leaders, technology companies, and experts – along with the knowledge and insights gained during her Harvard Fellowships – are helping to increase the profile, ambition and pace of the Call’s work and to bring on board new supporters and partners.

Over the last year, the New Zealand Government has strengthened the Secretariat, taking on new staff and facilitating secondments from the multistakeholder community to drive delivery of the Call’s global work programme and Leaders’ priorities.

The New Zealand Government has also continued to invest in domestic policies and programmes that respond to the tragic events of 15 March 2019 and the Royal Commission of Inquiry’s recommendations, and deliver against specific Call commitments – e.g. to understand and counter the drivers of terrorism and violent extremism in our society, to prohibit and prevent the dissemination of terrorist and violent extremist content online and to support frameworks for ethical media reporting on attacks. In this domestic work, there is a strong emphasis on transparency and human rights, and involving researchers, civil society and representatives of impacted communities.

Examples from the last year include:

- He Whenua Taurikura 2022 – the second counter-terrorism hui (meeting) – with a focus on all-of-society approaches to prevent terrorism and violent extremism in Aotearoa New Zealand. Almost 300 people attended from academia, communities, civil society, central and local government.
- The Department of Prime Minister and Cabinet delivered just over NZD $670,000 in funding to community and civil society organisations for initiatives to prevent and counter violent extremism in New Zealand.
- The Department of Internal Affairs released its second violent extremism transparency report, applying Tech Against Terrorism’s guidelines and providing insights into how the Department is applying its new regulatory take-down notice system, and the outcomes of that approach.
Where have you faced challenges?

The past 12 months have been characterised by a challenging economic backdrop, an evolving regulatory landscape, and the deployment of AI foundation models that are changing the nature of online services. These challenges have affected industry content moderation capacities, policies, and practices, which has at times resulted in cascading effects for communities harmed by or subjected to TVEC.

AI foundation models present both an opportunity and a challenge for The Christchurch Call in its work to eliminate TVEC. Work continues to better understand what is possible to proactively address risks. Similarly, The Call remains alert to the ways violent extremists and terrorists are beginning to use and exploit these technologies, and has stood up an emerging technology workstream as a result.

Smaller online service providers (OSPs) are particularly vulnerable to these challenges. A shortage of awareness and capacity to engage with the now-wide range of resources available to OSPs and an inability to participate in effective crisis response mechanisms remains a challenge. We expect these challenges to continue as governments and communities grapple with the many difficult questions and considerations that need to be addressed to ensure human rights are protected online. Violent extremists continue to exploit online environments.

Where does the Call Community need to focus effort in the year ahead?

The Christchurch Call acknowledges that new and emergent technologies continue to intersect at pace and depth across its focus areas. Remaining agile to these developments will be important to the Call’s future ability to lead its multistakeholder understanding of the opportunities and challenges this will present. The Call Community’s focus areas in the year ahead are therefore building on the work already completed or underway to advance these initiatives. This includes increasing our understanding of the implication of AI foundation models. We also want to engage with youth in more depth, assist small OSPs to further their crisis response capabilities, continue our work on understanding algorithmic influence, as well as sustaining and expanding the Call community. These focus areas have been identified through detailed Call community consultation and feedback.

Google

This statement was provided by Google

Overall, where has the Christchurch Call made progress in the last 12 months?
We commend the governments of France and New Zealand for continuing to progress the important work of the Call. In particular, we support the Call’s commitment to multistakeholderism, as evidenced by the Call’s progress in holding community dialogue sessions on important policy issues such as government transparency.

What have been your particular contributions/successes?
Jigsaw, in collaboration with Tech Against Terrorism (TAT), and with support from the Global Internet Forum to Counter Terrorism (GIFCT), is developing a free and open-source tool that aims to help smaller platforms address incoming flags of known terrorist and violent extremist content. We are currently working with a select group of member platforms of both TAT and GIFCT, and plan to make testing available to more platforms in the coming months.

In addition, and in line with the Call’s workstream on transparency, we have rolled out a new Transparency Center at transparency.google where people can easily access information about our
policies on a product-by-product basis, find our reporting and appeals tools, discover our Transparency Reports and learn more about our policy development process.

**Where have you faced challenges?**
Following the devastating attacks on civilians in Israel, our teams have removed thousands of harmful videos, and our systems continue to connect people with high-quality news and information. Our teams are working around the clock to monitor for harmful footage and remain vigilant to take action quickly if needed on videos, Shorts and livestreams.

Similarly, we are also detecting and removing content that violates our policies from other products, such as Ads, Shopping, and Play. Industry and multistakeholder groups like the Christchurch Call and GIFCT are key to ensuring appropriate coordination and sharing of best practices across industry and government, and helping the community adapt and streamline communications in response to attacks.

**Where does the Call Community need to focus effort in the year ahead?**
The Christchurch Community has provided an important avenue for engagement with civil society, including at the Call-hosted events and panels at the recent RightsCon Summit in Costa Rica. We strongly support the Call’s efforts to include communities affected by terrorist and violent extremist attacks in multistakeholder discussions to ensure their voices are heard.

We are also interested in leveraging the expertise of a multistakeholder forum to discuss and respond to the changing global regulatory landscape.

---

**Microsoft**

*This statement was provided by Microsoft*

**Overall, where has the Christchurch Call made progress in the last 12 months?**
Overall, the Christchurch Call has made progress in the last 12 months in continuing to build knowledge and relationships across the multistakeholder Community through a regular series of engagements and focused workstreams. The Call has also continued to evolve its focus to address new and emerging issues related to terrorist and violent extremist content: this is critical both to address the evolving threat and to recognize the intersectional nature of many online harms.

**What have been your particular contributions/successes?**
Microsoft has been pleased to contribute to the Christchurch Call Initiative on Algorithmic Outcomes, including through our $500,000 investment to support work to explore the role that privacy-enhancing technologies may be able to play in facilitating research on algorithmic outcomes. Through LinkedIn, we have participated in proof-of-concept testing, and we have also provided additional cloud computing support for OpenMined. Microsoft has also actively participated in a range of Christchurch Call workstreams over the last year, with a particular focus on new and emerging technologies. For example, we welcomed the opportunity to engage with the Christchurch Call Community on Bing’s new AI features. As a founding member of the Global Internet Forum to Counter Terrorism (GIFCT), we also continue support the GIFCT’s contribution to the Christchurch Call’s work, including on incident response and growing capability to address terrorist and violent extremist content across the tech ecosystem.

**Where does the Call Community need to focus effort in the year ahead?**
In the year ahead, Microsoft would welcome a continued focus on: - Continuing to grow the multistakeholder Call community and to leverage the Call’s model to help address emerging and related online harms; - Emerging technologies, including exploring the risks and safety opportunities
Global Community Engagement and Resilience Fund (GCERF)

This statement was provided by Global Community Engagement and Resilience Fund (GCERF)

Overall, where has the Christchurch Call made progress in the last 12 months?
Among the key issues listed for the 2023 Summit, we feel that most progress has been made by the Call through engaging its community on the growing threats of gender-based hatred and radicalisation to violence. The creation of a gender group, regular dedicated discussions, and a useful exchange of ideas and documents, have helped GCERF and the wider community advance our work on addressing this challenge across our global portfolio.

Progress has also been made in diversifying the civil society organisations providing input to the Call, reflected in GCERF’s membership which has the potential to mobilise the perspectives of over 200 local community-based organisations in 22 countries.

What have been your particular contributions/successes?
Since joining the Call a year ago, we led one CCAN session featuring diverse GCERF grantees, who spoke about challenges and solutions for increasing engagement with civil society on the ground. This engagement helped inform a new Global Counter Terrorism Forum (GCTF) Countering Violent Extremism Working Group initiative on ‘Funding and Enabling Community P/CVE’, launched on the margins of the UN General Assembly in New York in September. Here is an example of how the Call’s convening power can inform other global initiatives.

Considering promoting further information-sharing between GCERF and the Call’s community, GCERF has prioritised capacity-building amongst its grantees, and launched a new global P/CVE and Digital project, focusing on (i) how to use social media for P/CVE, (ii) critical thinking skills and digital literacy, and (iii) M&E for programs in digital P/CVE.

Where have you faced challenges?
We have faced no particular challenges regarding our emerging partnership with the Call, and again note the efforts of the Secretariat. As above, we believe there is potential for greater information-sharing, and perhaps joint initiatives and training, between GCERF’s and the Call’s communities, and we would welcome a discussion to develop a workplan.

Where does the Call Community need to focus effort in the year ahead?
Reinforcing a multi-stakeholder approach, and the continued expansion of relevant partnerships are key to achieving the Call’s Commitments. An increased focus on youth, in particular, is clearly needed. Following GCERF’s model, we would encourage consultation with civil society at every stage of the project design cycle.

European Commission

This statement was provided by the European Commission
Overall, where has the Christchurch Call made progress in the last 12 months?
In the last 12 months, the Christchurch Call and the EU have made progresses in addressing the potential negative effects of algorithmic amplification on online radicalisation. We welcome that the Christchurch Call has also made considerable progress in expanding the Call community to include broader representation from civil society and industry, reflecting the changes in the online threat landscape. We also welcome the progress made on protection of youth by setting up the Youth Reference group to investigate threats posed to children online that can lead to violent extremism.

What have been your particular contributions/successes?
The EU Internet Forum (EUIF) has delivered work on reaching an empirical understanding of the potential malicious effects of algorithmic amplification of violent extremist and borderline content. It has also produced a Handbook on Borderline Content, with guidance based on input provided by EU Member States, civil society organisations, the GIFCT and tech companies to better identify and prevent the spread of this content that is legal but harmful and can lead to radicalisation.

EUIF has also revised and further developed its EU Crisis Protocol, a voluntary mechanism that allows EU Member States and online platforms to respond rapidly and in a coordinated manner to the dissemination of terrorist content online in the event of a terrorist attack. The revised EUCP, among others, enables better coordination with international partners and addresses issues around bystander footage in crisis response.

Taking a user-centric approach in content moderation and empowering, in particular, young users is key to prevent radicalisation online. The EUIF has delivered a Handbook on Countering the Misuse of Gaming-related Content & Spaces based on input from civil society organisations and companies. The handbook guides companies on how to empower their users and gaming communities to identify, report and counter harmful content on their platforms.

Lastly, the implementation of the Terrorist Content Online Regulation and the Digital Services Act (DSA) in the EU has set important standards and rules to ensure that terrorist content and other types of illegal content are swiftly removed by tech platforms.

Where have you faced challenges?
While the majority of online platforms aim to prevent the exploitation of their service by violent extremist and terrorist actors, some platforms are providing a safe haven for malicious actors. Preventing the dissemination of harmful and illegal content on non-cooperative and small platforms continues to be a challenge that is facing the Call community.

Where does the Call Community need to focus effort in the year ahead?
Rapid technological development and quick adaption by malicious actors means that despite the progress we have made in the Call community old challenges remain and new threats are emerging. We believe that more information, expertise and conclusive exchange is needed to better understand the extent of the threat posed by the malicious use of new technologies, such as generative AI, immersive reality and decentralisation to develop preventive measures and actions to tackle the threats these may pose. Moreover, we need to have more expertise and greater engagement from tech companies in the moderation of harmful and violent extremist content in relation to gender-based hate and violence. While progress has been achieved on issues surrounding algorithmic amplification, we need concrete follow-up actions and transparency on the functioning of their algorithmic systems. In addition, terrorist operated websites remain a challenge and greater engagement with internet infrastructure providers is needed to address this threat effectively.

Tech Against Terrorism
**Overall, where has the Christchurch Call made progress in the last 12 months?**

In 2022/23 Tech Against Terrorism increased its direct engagement with the Call by sharing intelligence both in the course of monitoring terrorist content generally and in response to crises, and by engagement on policy matters in the working groups. We greatly appreciate the efforts made by the Call to build its engagement with its partners whilst ensuring that this engagement is action-oriented. Tech Against Terrorism has particularly appreciated the launch and structure of the working groups, which from an early stage were concerned with creating practical guidance for the Call Community. The structure of the working groups greatly reflects the concerns and priorities identified by the Community in previous years. Tech Against Terrorism notes its alignment with the priority focus of the Call over the course of the last year, in particular around transparency reporting for government, improved coordination at times of crisis, and engagement with infrastructure providers. This priority incorporates issues highlighted by Tech Against Terrorism in the previous round of consultations.

**What have been your particular contributions/successes?**

In 2022/23 Tech Against Terrorism increased its direct engagement with the Call by sharing intelligence both in the course of monitoring terrorist content generally and in response to crises, and by engagement on policy matters in the working groups. We greatly appreciate the efforts made by the Call to build its engagement with its partners whilst ensuring that this engagement is action-oriented. Tech Against Terrorism has particularly appreciated the launch and structure of the working groups, which from an early stage were concerned with creating practical guidance for the Call Community. The structure of the working groups greatly reflects the concerns and priorities identified by the Community in previous years. Tech Against Terrorism notes its alignment with the priority focus of the Call over the course of the last year, in particular around transparency reporting for government, improved coordination at times of crisis, and engagement with infrastructure providers. This priority incorporates issues highlighted by Tech Against Terrorism in the previous round of consultations.

In 2022/23 Tech Against Terrorism increased its direct engagement with the Call by sharing intelligence both in the course of monitoring terrorist content generally and in response to crises, and by engagement on policy matters in the working groups. We greatly appreciate the efforts made by the Call to build its engagement with its partners whilst ensuring that this engagement is action-oriented. Tech Against Terrorism notes its alignment with the priority focus of the Call over the course of the last year, in particular around transparency reporting for government, improved coordination at times of crisis, and engagement with infrastructure providers. This priority incorporates issues highlighted by Tech Against Terrorism in the previous round of consultations.

**Transparency reporting.**

We contributed significantly to the working group on transparency reporting and subsequently to the development of the Community Advice on Government Transparency. Tech Against Terrorism assumed a prominent speaker role a speaker role during the first meeting and actively participated throughout all the working group sessions, including the provision of feedback on the draft Community Guidance. Our contribution to the Call’s transparency efforts is further in evidence in the Community Guidance on Government Transparency, which are based on Tech Against Terrorism’s Guidelines on Transparency Reporting on Online Counterterrorism Efforts for Governments. Content identification and TUI monitoring. We also support the Call in the broader identification of TVE content online by sharing the results of our monitoring and investigations. In doing so, Tech Against Terrorism proactively assists in identifying material related to the Christchurch attack and alerting it to tech companies. A recent example of this was Tech Against Terrorism’s support in locating copies of the videos of the pseudo-documentary “The Three-Faced Terrorist”.

**Where does the Call Community need to focus effort in the year ahead?**

We recommend the following issues to be prioritised: Terrorist-Operated Websites (TOWs) remain a major threat by allowing terrorists and violent extremists to share their propaganda both widely and in a stable environment. There are a number of barriers and complications currently impeding action against TOWs including jurisdictional gaps between governments, within governments, and between governments and tech companies clouding the issue of who should lead, request, and coordinate action on TOWs. The Call should lead on a global coordinated strategy to counter TOWs, and support Tech Against Terrorism to drive engagement with infrastructure providers to ensure the effective disruption of TOWs. Given its transformative potential, it is imperative that tech platforms begin to design policies and processes to prevent terrorist exploitation of generative AI, whilst ensuring the safeguarding of human rights and freedom of expression online. Governments also have a responsibility to regulate and set the parameters for legal use of generative AI, given the real risk that TVEs will abuse emerging technologies. In addition to mitigating against the risks of TVE abuse of
AI technology, it is also critical to discuss the opportunities AI provides for improving responses to the threat in the areas of content moderation, content detection, creating classifiers, supporting behavioural analysis, and providing more rapid responses in crisis situations. The Call should continue its investigation into the risks and benefits of AI started by the working group on Emerging Technologies. The Call should continue its work to coordinate responses at times of crisis, and to ensure in particular the monitoring of the online footprint of TVE attacks as well as the provisions of post-crisis debriefing. Tech Against Terrorism will continue to support the Call’s crisis coordination efforts by alerting material related to crisis events via the with its reach of 120 platforms.

Cyber Threats Research Centre *(CCAN member)*

This statement is provided by the Cyber Threats Research Centre (CYTREC), Swansea University. CYTREC is an interdisciplinary research centre exploring a range of online threats from terrorism, extremism, and cybercrime, to child sexual exploitation online and grooming based at Swansea University. Members have backgrounds in law criminology, political science, linguistics, psychology, and computer science. CYTREC is a CCAN member.

**Overall, where has the Christchurch Call made progress in the last 12 months?**
The Christchurch Call Initiative on Algorithmic Outcomes (CCIAO) with the aim to develop new tools to empower research in this area, has the potential to facilitate real impacts in this area. Facilitating engagement and discussion across all sectors in the Call Community.

**What have been your particular contributions/successes?**
We have participated in the Algorithmic Oversight and Transparency Workstreams. One of our members is currently a Co-Chair of the CCAN.

Some recent and most relevant work has included studies on the role of social media recommendation algorithms and potential links to extremist content. An ongoing project on recommendation systems involves conducting a scoping review of all of the available literature on the amplification of potentially harmful online content. In addition, we have been working on issues of moderating so-called ‘borderline content’ online whilst respecting fundamental values. And, on defining ways that extremists circumvent AI content moderation practices by manipulating platform affordances.

CYTREC is part of a consortium of partners led by Tech Against Terrorism working to boost the resilience of small and micro platforms against terrorist exploitation and helping them to comply with their responsibilities under the EU’s Terrorist Content Online Regulation as part of the Tech Against Terrorism Europe project. We are part of the Researcher, Security, Safety and Resilience (REASSURE) project, which is about enhancing the wellbeing and safety of online extremism and terrorism researchers through active collaboration with the community of scholars in this field. The project is working towards producing a Charter for Researcher Ethics and Safety.

We lead the ‘Seeing Eye to Eye: Developing Sustainable Multistakeholder Communities’ project working collaboratively with stakeholder participants to develop an empirical assessment of barriers, benefits, institutional norms, trust and transparency when working in multistakeholder environments. Other projects have looked at the way different stakeholders frame Kenya’s extremist problem and the available spaces for preventing and countering violent extremism interventions. And, hate speech as a driver of social media engagement using a case study from Turkey.

**Where does the Call Community need to focus effort in the year ahead?**
- The development of the governance and ethics frameworks in collaboration with the Call Community to take the CCIAO to the next stage.
- Building and sustaining the Call Community. Including building on existing relationships to continue to enhance engagement across sectors in the Community.
- Identifying and measuring the impact of the work of the Call and the Community.

Federation of Islamic Associations of New Zealand (FIANZ) (CCAN member)

This statement is provided by Federation of Islamic Associations of New Zealand (FIANZ) Abdur Razzaq Chairperson, FIANZ Royal Commission Submission and Follow-Up

Overall, where has the Christchurch Call made progress in the last 12 months?
As the umbrella national Muslim organisation of Aotearoa New Zealand, our responsibilities post-March 15 have extended well beyond normal community development programmes to stridently focus on ensuring the safety of our small faith but ethnically diverse community. The implementation of the Royal Commission Recommendations on effective PCVE is a key priority for us. It is in this context we joined the Christchurch Call. On-line extremism and hate-mongering are not just virtual threats for us but their real-world transition and trajectory is tangible as evidenced by the terrorism we experienced at Masjid An-Nur and Linwood Masjid with 51 shuhada. Whilst the March 15 terror was aimed at our faith, it could have well been other vulnerable communities. In this context, From FIANZ’s perspective, the use-value of the Christchurch Call over the last 12 months have been threefold.

1. Effective lobbying at a global stage with governments, multi-lateral agencies and platform companies
2. Formative strategies adopted to counter on-line threats as well as the nuanced mis- and dis-information with particular focus on algorithmic solutions.
3. Seeking accountability of countries and companies where issues have emerged and which members consider there is a need to challenge

The three need-value metrics we have used to measure the efficacy of the Christchurch Call have been:

1. Availability of Information (Everyone has been given access to and the necessary time-lag to digest information.)
2. Opportunity to Contribute (The work streams have been carefully thought out and fosters an inclusive and transparent approach, despite the boundaries of geography and time)
3. Democratic Oversight (Everyone has an opportunity to raise critical issues on matters of governance, choice of agenda topics, liaison with external stakeholders and such. This healthy introspection engenders both respect and trust in the Call initiative)

What have been your particular contributions/successes?
After March 15, our Muslim community experienced an endless series of consultations and engagements, particularly with government agencies. This has resulted in an aversion to ‘talk-fest’
sessions. Christchurch Call however has been quite different in this respect. By participating in the Call, we have been able to realign our community’s Strategic On-Line Safety (SOS) programme. For the last 7 months we have commenced implementation with on-going formative evaluation to guide improvement.

The SOS covers the continuum:

1. A small team to monitor on-line harm and when necessary inform NZ authorities;
2. Develop on-line counter-narrative to address ‘baited escalation’ of extremism;
3. Help our community, particularly the youth, to discern online mis- and dis-information through education programmes (NZ Censor Office is also assisting);
4. A number of consultation platforms with key government agencies (NZ Police, NZSIS, DIA, Justice, Education and others) to address emerging issues;
5. Regular community programmes which also includes nexus to on-line harm/extremism. Many of these sub-programmes have benefited from our participation in Christchurch Call.

Where have you faced challenges?
The Call members, with their wealth of experience and expertise in different fields, have provided a lot of information and practical suggestions. In this context, the key challenge we have faced is translating the lessons learned at the global level to practical initiatives with measurable outcomes at the local level.

Where does the Call Community need to focus effort in the year ahead?
Continue the current focus on such areas algorithms, transparency and others and include: a) AI in countering on-line extremism b) Local Communities, Governments and Platform Agencies to collaborate to develop and promote education programmes, particularly for our children and youth.

Global Internet Forum to Counter Terrorism (GIFCT)

This update was provided by the Global Internet Forum to Counter Terrorism (GIFCT)

The Global Internet Forum to Counter Terrorism (GIFCT) brings together the technology industry, government, civil society, and academia to foster collaboration and information-sharing to counter terrorist and violent extremist activity online. GIFCT’s mission since 2017 has been to prevent terrorists and violent extremists from exploiting digital platforms. The Christchurch Call to Action acted as the impetus for GIFCT’s founding tech company members to transform GIFCT into a 501c3 non-profit and for them to develop a 9-Point Action Plan that set out concrete steps the technology industry would take to address the abuse of technology to spread terrorist content, including both individual and collaborative commitments.

GIFCT has continued to participate in and build upon the Christchurch Call’s work streams and thematic focuses in the years since its inception. GIFCT is submitting this update to the Christchurch Call to outline the work that has been carried out over the last twelve months, what GIFCT has in store in the future, and how we plan to further our joint efforts to counter terrorism and violent extremism online in the years to come.

Overall, where has the Christchurch Call made progress in the last 12 months?
The Christchurch Call has continued to evolve and expand its focus and remit into new topics, in line with what the Community sees as an evolving threat, such as looking more concretely at violent extremism as it affects gender, younger online audiences, and smaller platform needs. GIFCT has also noticed a greater number of NGOs joining the Call and it would be useful to review the expectations of current government, tech, and NGO members to remind members how they are meant to contribute or highlight progress against commitments throughout the year.

**What have been your particular contributions/successes?**

Since the launch of the Christchurch Call 9-Point Action Plan in 2019, GIFCT has strengthened its joint commitments made with GIFCT Member companies. Over the past year, line with themes of interest put forward by the Christchurch call, GIFCT has:

1. **Expanded its Membership:**
   - GIFCT has brought on 10 new members to GIFCT since the 2022 UN General Assembly.

2. **Further Integrated Human Rights Due Diligences in its Work:**
   - GIFCT has added reviews of its Members’ human rights commitments in line with the UNGPs. GIFCT collaborates with Business for Social Responsibility (BSR) in providing human rights engagements with aspiring Members.
   - GIFCT has further embedded human rights due diligence in its Incident Response Protocols based on input from the GIFCT Incident Response Working Group, including hosting multistakeholder debrief sessions after any Content Incident Protocol.

3. **Sustained and Added Nuance into GIFCT’s Incident Response Framework:**
   - GIFCT and members have initiated communications to share situational awareness and information in response to over 306 terrorist or mass violence events and significant online terrorist developments in 44 countries across six continents.
   - GIFCT and Meta hosted a member company Hackathon exploring ways to enhance our current processes and procedures for situational awareness and information-sharing during a developing terrorist or mass violent event in order to strengthen our collective readiness and agility to respond to the event’s online dynamics. This effort feeds directly into our work to constantly review and update our Incident Response Framework.
   - Since the GIFCT Incident Response Framework (IRF) was put in place in 2019, GIFCT has activated the highest level of the IRF—the Content Incident Protocol—six times in response to the perpetrators of mass violent attacks livestreaming their violence. This includes the CIP activation in Louisville, Kentucky in April, 2023.

4. **Advanced Understanding of Threats and Opportunities of Joint Technologies:**
   - GIFCT has steadily increased the amount and type of content signals shared through the Hash Sharing Database, containing approximately 370,000 unique and distinct items relating to approximately 280,000 visually distinct images, 90,000 visually distinct videos, and 200 textually distinct items related to PDFs.
   - GIFCT published a rigorous technical trial on Combining Behavioural Signals to Surface Terrorist and Violent Extremist Content Online to test layered threat detection models for proactively surfacing awareness of ongoing terrorist and violent extremist attacks.
GIFCT progressed on a series of joint technical solutions. As an example, Faculty AI and GIFCT have developed a model that facilitates small platforms to access terrorism classification models at no cost.

Using a multistakeholder approach to identifying and analyzing future threats related to technological advancements, GIFCT’s Year 3 Red Teaming Working Group launched a series of expert insights at UNGA on

1. Social Media and its Impact on Terrorism and Violent Extremism in the Next 2-5 Years,
2. Risks and Challenges in Online Communities for 3D-Printed Firearms Among Extremists and Terrorists, and
3. Considerations of the Impacts of Generative AI on Online Terrorism and Extremism. In addition, further expert insights are planned focusing on end-to-end encryption and federated web technologies.

As part of the planned Taxonomy Expansion to the GIFCT Hash Sharing Database, we have developed the capability to include URL’s into the Hash-Sharing Database to support members in addressing outlinking to terrorist content from their platforms.

GIFCT has also ensured support to ongoing technical efforts being progressed by partners at Meta, Jigsaw, and Tech Against Terrorism.

Developed Applicable Insights on Borderline Content:

As a key theme brought up by the Christchurch Call Community and the EU Internet Forum, GIFCT developed a report on Borderline Content. This GIFCT contribution to the EU Internet Forum’s discussions of borderline content provides guidance on the term itself and a nuanced discussion around the relevant online policies and practices used by GIFCT member companies for borderline content.

Furthered Studies of Gender and Online Extremism:

Advancing a salient topic put forth by the Christchurch Call, GIFCT has worked with its academic network, the Global Network on Extremism and Technology (GNET), producing a series of expert insights on Gender and Online Violent Extremism. These pieces focus on how terrorist and violent extremist groups target and/or recruit based on gender and identity across geographically diverse settings. The series includes a wide array of analyses of how misogyny and attacks on the LGBTQIAP+ community have manifested in recent years.

Furthered Understanding on the Role of Algorithms in Relation to Processes of Radicalization:

GIFCT’s Year 2 GIFCT Working Groups (2022) produced a series of outputs analyzing the role of algorithmic amplification within processes of radicalization. While this is from last year, it speaks to the Call’s continued interest in algorithmic processes. CCAN contributed largely to these efforts

Methodologies to Evaluate Content Sharing Algorithms and Processes (p.115 - 151) laid out the process for querying an algorithmic system and determined if it was inadvertently surfacing terrorist or violent extremist material.

Dr. Jazz Rowa’s report on the Contextuality of Lone Wolf Algorithms found that the relationship between the user, the outside world, and algorithmic systems were inextricably linked, meaning the agency of the user had to be taken into account. The report also outlined
clear pathways for working within algorithmic systems to develop safety-by-design in defining what could be added or taken away from a system to ensure a safer environment online.

- An empirical review (p. 162 - 191) on all main available academic reviews of algorithmic amplification in relation to hate based and violent extremist content found that there was a lack in clear definitional frameworks for terms such as “extremism”, with implications for the debate around borderline content and what constituted illegal or policy-violating content.

**Where have you faced challenges?**

The global political environment will continue to affect how tech-focussed and multistakeholder initiatives tackle terrorism and violent extremism, as governments do not have agreed upon definitions or designation frameworks. There is also a mainstreaming of violent and hate-based organizations in certain parts of the world, while there continues to be a decrease in funding to CSOs and NGOs that traditionally upheld PVE/CVE work. As violent extremist groups and sympathizers become increasingly decentralized, this sector will have to redefine its scope in line with its core mission in order to remain effective and relevant.

**Where does the Call Community need to focus effort in the year ahead?**

- GIFCT would like to ensure that its partnership with the Christchurch Call Community continues with clarity on thematic focuses and an understanding of how best to engage the Community and CCAN in progressing its shared goals. What does success look like for the Christchurch Call Community? GIFCT looks forward to progressing its work in alignment with the Call.

---

**Institute for Strategic Dialogue (CCAN member)**

*This statement was provided by the Institute for Strategic Dialogue (ISD)*

**Overall, where has the Christchurch Call made progress in the last 12 months?**

From our perspective, while not involved across all of them, good progress has been made across the various different working groups and key areas identified, and the focus on tangible commitments and outputs has been encouraging. The Call Secretariat, working group, and CCAN leads have all been proactive in engaging, and well organised, which is not always the case with international, multistakeholder efforts and initiatives.

**What have been your particular contributions/successes?**

Over the past year ISD has participated in the CCAN and the New Technologies and Gender working groups. We regularly participate in calls across these groups, as well as the central network calls, share relevant ISD research, and aim to ensure that our work in other areas/via other projects (e.g. our Digital Policy Lab) both complement and do not conflict with or duplicate ongoing efforts under the Call or other related international and multistakeholder initiatives.

Over the last 6 months, ISD and the Call have collaborated particularly closely on the topic of gender based online violence/misogyny and its link to radicalisation and TVEC. Within the framework of ISD’s Digital Policy Lab, we held an online working group series including members of the CCAN, online service providers, and representatives from ministries and regulators from DPL member countries. The discussions have covered the role of social media platforms in the proliferation of GBOV and explored potential actions and responses that platforms can and should take to address OGBV. ISD has released a policy brief summarising these discussions, and highlighting mitigation actions platforms can take. Members of the Call Secretariat team have been particularly helpful in
supporting and contributing to this effort, and encouraging others in the network to also actively participate.

Where have you faced challenges?
Resourcing / capacity to engage: With a relatively small digital policy team it can sometimes be difficult to engage on a voluntary basis as fully as we’d like in all areas of interest, so we have to be more selective in terms of the working groups we can participate in, events we can attend (e.g. RightsCon) etc. This also applies to other multistakeholder initiatives (e.g. GIFCT, Global Partnership), so any opportunities to consolidate, streamline and/or deconflict different initiatives is welcome.

We completely recognise that scheduling calls at times when everyone can participate across time-zones is essentially impossible/unavoidable, and others will face the same issues, but it does mean that sometimes we are not able to participate in every monthly, working groups or CCAN call.

- Where does the Call Community need to focus effort in the year ahead?
  Continue the progress made across the various working groups to ensure the next steps are followed up on/achieved over the next annual cycle as far as possible.
  - Continue to develop metrics/benchmarks to quantitatively assess the impact of the Call and its industry members efforts to prevent and combat TVEC online.
  - Ensuring incoming regulation across Call jurisdictions is both effective at addressing harms stemming from online terrorism, violent extremism and hate speech, while also upholding and protecting human rights.
  - Ensuring that increased transparency from companies and data access for regulators and researchers helps the community to develop a better understanding of current gaps in the evidence base on TVEC online, as well as the impact of platform systems/design, the effectiveness of platform mitigations, and the level of resources companies dedicate to these issues.
  - Supporting other governments around the world that are considering new internet legislation to ensure compliance with international human rights standards when attempting to address TVEC online.

InternetNZ (CCAN member)

This statement was provided by InternetNZ

Overall, where has the Christchurch Call made progress in the last 12 months?
The various working groups have done some encouraging work, particularly in the areas of Online Gender Based Violence (OGBV) including the work conducted with ISD, and Government Transparency. There are some promising steps in the areas of Youth in relation to TVEC, supporting smaller platforms, and addressing new technologies such as AI and LLM, as well as a commitment to continue work assessing algorithmic amplification. The (re-)establishment of the Free, Open and Secure Internet working group is a good step as well, as its role will help ensure human rights are not undermined in the Call’s CVE efforts.

In sum, there has been some really positive issue-based work that supplements more general efforts to broaden the Call’s membership and points to some useful areas of development in the future.

What have been your particular contributions/successes?
InternetNZ Senior Policy Advisor Michael Daubs has been a regular participant in CCAN meetings and a member of the Free, Open and Secure Internet working group. In addition, he has contributed to
the planning meeting for the Hui Taumata, an attempt to have the first in-person CCAN meeting in Christchurch in May of next year, both to mark the 5th anniversary of the 2019 terrorist attack there, and to have an "unconference" to develop community-based approaches to CVE that can inform future Call efforts. InternetNZ also provides structural support for CCAN.

Where have you faced challenges?
The biggest challenges lie in the area of capacity; both limited time and resources affect our contributions to both CCAN and the Call.

Where does the Call Community need to focus effort in the year ahead?
Continuing the focus on government transparency and algorithmic outcomes should be a focus, as should emerging technologies, particularly the potential uses of generative AI in TVEC. In addition, a focus on radicalisation pathways beyond algorithmic recommender systems deserve some consideration.

Internet Society (CCAN member)

This statement is provided by Internet Society

Overall, where has the Christchurch Call made progress in the last 12 months?
In the past year, the Christchurch Call sought to strengthen the participation of the technical community in its work. It reconvened the Free, Open, and Secure Internet Expert Group to advance a shared understanding of the Call’s pledge to uphold an open and secure Internet among its supporters. CCAN secured greater involvement for civil society and technical experts at the GIFCT, helping to ensure that industry measures to foster a healthy online environment are informed by a broad range of perspectives.

What have been your particular contributions/successes?
As a member of CCAN, the Internet Society endeavors to ensure that efforts to mitigate the spread of terrorist and violent extremist content online safeguard the Internet of opportunity. In March, we testified at the US Senate Hearing on Platform Accountability, emphasizing the continuing relevance of safe harbor provisions in fostering innovation, collaboration, and knowledge exchange online. Likewise, we took part in the UN Security Council Counter-Terrorism Committee’s Executive Directorate consultations to highlight the need for targeted and fit-for-purpose principles to minimize the risk of unintended consequences to Internet users.

Through our Policymakers’ Fellowship track at Internet Engineering Task Force meetings, we engage with governments to build their technical literacy, and in May, we collaborated with the governments of Canada and the UK on a capacity-building workshop at the Conference on Countering Radicalization to Violence.

At Internet governance fora, we continue to highlight the Christchurch Call as a prime example of a strong multistakeholder initiative to address emerging issues.

Where does the Call Community need to focus effort in the year ahead?
As the Call develops more mechanisms to combat terrorist and extremist activity online, it needs to encourage its supporters to more consistently assess the impact that their policies may have on the free, open and secure Internet, using frameworks such as the Internet Impact Assessment Toolkit. Doing so will ensure that actions to progress the Call’s commitments do not adversely affect the value of this global resource to economy, society and for future generations.
United Nations Counter-Terrorism Committee Executive (UN CTED)

This statement was provided by United Nations Counter-Terrorism Committee Executive Directorate (CTED)

Overall, where has the Christchurch Call made progress in the last 12 months?
The Christchurch Call has made progress in increasing transparency, bringing in senior leaders and political buy-in at the highest levels of many of its members, and in focusing on technical (and sensitive) issues such as algorithms.

What have been your particular contributions/successes?
CTED supported the United Nations Security Council’s Counter-Terrorism Committee (the Committee) to hold a Special Meeting on countering the use of new and emerging technologies for terrorist purposes in Mumbai and New Delhi, India on 28 and 29 October 2022. In preparation for the Special Meeting, CTED hosted a series of technical meetings on ICT topics in late-September and early-October 2022. The Christchurch Call participated in the technical meetings. During the Special Meeting, the Committee focused on three significant areas which emerging technologies are experiencing rapid development, growing use by Member States for security and CT purposes, and an increasing threat of use for terrorist purposes.

One focus area was the use of Information and Communication Technology (ICT), namely the Internet, social media, and related digital and online spaces, for terrorist purposes. The Christchurch Call was invited to speak during the Special Meeting. As the primary output from the Special Meeting, the members of the Committee formally adopted the “Delhi Declaration” outcome document on 29 October 2022. The Delhi Declaration recognizes a number of key concerns and provides guidance to Member States and stakeholders in countering the use of new and emerging technologies, including ICT, for terrorist purposes and called for the development of recommendations and non-binding guiding principles (NBGPs).

In its work to develop the ICT guiding principles, CTED conducted 9 consultative sessions with key stakeholders, including a number of private sector and civil society Christchurch Call Community entities. CTED continues to work with the Committee on the development and eventual adoption and publication of the ICT guiding principles and related recommendations. CTED will additionally organize an open briefing of the Counter-Terrorism Committee on preventing and countering terrorist narratives online and offline, as per Security Council resolution 2354 (2017).

Where have you faced challenges?
While CTED has been able to progress in its ICT-related assessment work with UN Member States and increase its efforts to raise ICT-related issues to the attention of the CTC and Security Council, CTED recognizes that there is no global consensus regarding some of the identified threats and possible ways to address them, particularly as regards the role of technology companies, Governments, civil society organizations and other stakeholders.

There are also differing levels of compliance with the rule of law and international human rights law, as well as respect for fundamental freedoms by governments (and technology companies) in countering terrorism online and through cyber means. This lack of political agreement on ways forward can slow some of the UN’s consensus-driven normative and policy-related work.

Where does the Call Community need to focus effort in the year ahead?
In the year ahead, the Call Community could work more closely with the United Nations, including CTED and other UN bodies with a remit for preventing and countering terrorism. For example, participation of Christchurch Call leaders in CTED events and briefings to the UN Security Council and
its Counter-Terrorism Committee could encourage increased understanding and support from Christchurch Call Member States, as well as those not yet part of the Call Community. Of note, Security Council resolution 2617 (2021), which first mentioned “other emerging technologies” in connection with CTED’s work on countering use of the internet and other ICTs, also notes the Christchurch Call to Action.

CTED would encourage even stronger engagement between the Christchurch Call and Tech Against Terrorism to ensure that small platforms receive the support and training necessary to address the misuse of their tools for terrorist and violent extremist purposes.

The Call Community could additionally focus on supporting the creation and availability of digital literacy programmes (including digital media literacy and AI literacy) for government authorities, civil society organizations, media representatives/journalists, and all levels of society to address mis/disinformation and to help identify and prevent the distribution of terrorist and violent extremist-related content. Coordination with NPOs/companies such as Mythos Labs could facilitate this. The Call Community entities and Christchurch Call members and partners should continue efforts to collectively discuss and attempt to develop consensus on what the responsible use of artificial intelligence, algorithms, and strategic communications might look like.

Global Project Against Hate and Extremism (CCAN member)

This statement was provided by Wendy Via, President and Co-Founder, Global Project Against Hate & Extremism (GPAHE)

Overall, where has the Christchurch Call made progress in the last 12 months?
Overall, the Christchurch Call has made progress in the last 12 months in the very difficult area of addressing the accountability of Call members to uphold their commitments. Given the spectrum of members, this is an understandable challenge and one where we’re glad to see attempts at navigation.

What have been your particular contributions/successes?
At GPAHE, we believe that fully participating in the strategic planning, meetings and discussions of CCAN is the most important thing we can do. The partnerships and fellowship and ensuing trust within CCAN are critical to its success.

As an organization, our work is specifically about preventing far-right extremism and associated violence, especially that which is inspired and organized in online spaces. GPAHE has taken forward the work of the Call by collaborating with several companies on the development of their dangerous organizations and individuals policies and providing expert advice to governments developing regulations for the tech industry. We’ve also exposed violative content on the platforms, for example the “Great Replacement” conspiracy theory. After the Buffalo shooting, the companies were eager to show the steps they were now taking to address the problem even though there had been multiple shootings in the three years since the Christchurch shooting, also inspired by replacement theory, and the genesis of the Call. There is still much work to do in this area.

GPAHE has had considerable impact in the field. Among other efforts, we exposed the online activities of the largest international network of violent white supremacists and were able to work
with companies to deplatform these groups. We also documented how some political figures use the platforms for hate and dehumanizing speech and disinformation, threatening democracy and inspiring violence. We’ve formed a global alliance of more than 70 human rights groups to address this challenge. And more recently, we released our Global Extremist Symbols Database, used by tech companies and rights advocates around the world.

We are keen to share our expertise and work with the Community in support of the Call.

Where does the Call Community need to focus effort in the year ahead?
Going forward, the Call needs to best identify where CCAN can be the most helpful in achieving the commitments. One concrete step toward adding more value would be to clarify the remit of the advisory network, call on the expertise of its members, and empower CCAN to put forth actionable plans for ensuring accountability. Accountability of the Call members should be a focus in the coming year, especially since so many countries and billions of people will be participating in elections.
2023 Issues Brief

This *Issues Brief* presents views from across the Christchurch Call’s global multistakeholder Community about the progress we have made, individually, collectively, and in groups to deliver the Call’s 25 Commitments. The *Issues Brief* looks at our achievements, areas where we need to adjust or increase effort, priority issues for the future, and difficult questions that require discussion or strategic guidance.
Introduction ........................................................................................................................................... 2
Purpose ................................................................................................................................................... 2
How the Issues Brief was developed .................................................................................................... 2
How it fits with our commitments and work programmes ...................................................................... 2
Assessment of key issues ....................................................................................................................... 3
  1. Implications of new technology, including foundation AI models. .............................................. 3
  2. Assisting smaller OSPs to eliminate TVEC online and respond to crisis incidents. ............ 5
  3. Deepening the evidence base on gender-based hatred & radicalisation to violence; ...... 7
  4. Algorithmic outcomes, young people, & pathways to radicalization. ....................................... 9
  5. Christchurch Call Initiative on Algorithmic Outcomes (CCIAO) ................................................ 10
  6. Building, sustaining, and expanding the Call Community ............................................................ 12
  7. Multistakeholder approaches to related issues ............................................................................. 14
Other issues raised. .................................................................................................................................. 15
Conclusions ............................................................................................................................................ 16
Introduction

Purpose

The *Issues Brief* will be used as a basis for the Agenda for the 2023 Leaders’ Summit and the Summit co-Chairs Joint Statement, and discussion topics for our subsequent whole-of-Community Meeting. The process of developing this brief will allow us to collaborate between and across sectors as a Call Community, enhancing our dialogue about important issues.

How the *Issues Brief* was developed

The *Issues Brief* is based on input from our working group leads, from the Community through our 2023 Survey, through dedicated Town Hall meetings, and other written input.

The *Issues Brief* is representative of diverse perspectives from our Community. It is not necessarily the consensus view of every supporter, partner organisation, or civil society network member. The 2023 *Snapshot Report* provides information and perspectives from individual Community members and organisations.

How it fits with our commitments and work programmes

Call supporters fulfil their commitments to eliminate terrorist and violent extremist content (TVEC) through joint and individual efforts, and have undertaken to do so transparently, while upholding human rights and a free, open, secure internet. The 25 *Call Commitments* remain the enduring foundation of our activities.

In 2021, we launched *work plans* to help align our efforts in four areas where we saw particular importance and shared interest in working together:

1. Building a diverse, enduring, and effective Call Community,
2. Understanding user journeys and algorithmic impacts and creating a new generation of community-led online [positive]? interventions,
3. Delivering meaningful transparency about government and industry actions, and
4. Responding effectively to terrorist and violent extremist incidents whilst putting in place meaningful human rights and due process mechanisms.

For the 2022 Summit, the Community worked over several months to identify and understand new challenges, or issues affecting delivery. This helped develop an agenda and guiding questions to enable Leaders to address the most important strategic challenges and make a meaningful difference through their leadership.

In 2023, we face a different economic backdrop, a changed regulatory landscape, and the deployment of new AI models that are changing the nature of online services. Communities around the world continue to grapple with difficult questions over content moderation, the protection of human rights online, radicalisation to violence, and harm from TVEC.

The following issues formed the starting point for our 2023 pre-Summit engagement as a Call Community and have helped us to identify key actions and discussion questions for Leaders:

- The implications of new technology, particularly foundation AI models;
- Assisting small online service providers (OSPs) to eliminate terrorist and violent extremist content online and respond to crisis incidents;
- The evidence base on gender-based hatred, acts of gender-based violence online and offline, and radicalization to violence;
- Youth, algorithmic outcomes, and pathways to radicalization;
- The Christchurch Call Initiative on Algorithmic Outcomes;
- Sustaining and expanding the Call Community;
- Multistakeholder approaches to related issues.

This Community engagement process is important to identify areas that require focus, prioritisation, resources, adjusted approaches, or ongoing attention. This will help the Community stay ahead of changes in the environment and deliver our long-term vision and commitments.

Assessment of key issues

This section sets out in detail the key issues the Community is working on, to provide a common understanding of their significance, what has been achieved, and where there are challenges and opportunities ahead. This in turn has helped to identify the smaller number of specific topics that may usefully be discussed among Leaders at the 2023 Summit.

1. Implications of new technology, including foundation AI models.

Why it’s important...

The nature of online services is evolving rapidly. A growing number of AI-driven tools and interfaces are embedded in the online services we use every day. This brings potential mechanisms for harm and exploitation, and new and potentially useful prevention and intervention strategies.

At the 2022 Summit, Leaders asked for guidance on how to support the adoption of new technologies, while addressing the impacts this may have for the elimination of TVEC, and the potential new forms of TVEC we may encounter and need to address.

The 2023 Summit is an opportunity for dialogue about new technology, particularly AI foundation models, the impacts and policy implications, and the views of communities and civil society.

This is one of the most complex problems confronting policymakers today. The Call offers a specific and practical lens through which to assess identifiable harms and risks, and frameworks for addressing them.

Recent progress and efforts by the Community

The Call Community launched a New Technology Working Group to build its understanding of the risks and opportunities at the nexus of new technologies and TVEC.

The New Technology Working Group has identified key risks associated with generative AI, immersive technologies, and the decentralised web, and is now developing advice on strategies to mitigate these risks.
The New Technology Working Group has met with the Call Community and individual supporters to understand their new products e.g. the new Bing AI tool, gathering input about possible concerns and the safety features being deployed to counter TVEC.

At RightsCon, the Call Community hosted a workshop inviting civil society to discuss the implications of new technology for TVEC and ways to address this.

The Global Internet Forum to Counter Terrorism (GIFCT) and Institute for Strategic Dialogue (ISD) are both developing expert outputs on new technology, including approaches to red-teaming and how generative AI could affect content harms.

**Difficult questions**

How should industry, government, and civil society align and prioritise their expectations and priorities for safety? How can they distinguish genuine problems from perceived or theoretical problems and address them with an appropriately calibrated level of precaution and foresight?

How can we balance the desire for aligned approaches to safety with the importance of openness to competition and innovation, and access and autonomy for users?

How can we ensure that safety interventions fitted onto these systems are consistent with a free, open, secure internet and don’t impede access and innovation for users around the world?

How should we address emergent harms that may arise unexpectedly as new products are commercialised and become widely available?

The blurring of content types and the lines between content, human and machine behaviour in a new technology environment make it increasingly difficult to establish taxonomies and define norms. How can we do so in a consistent way, and preserve our commitments to human rights, journalistic freedom, and freedom of expression?

Decentralisation of technology, tools, and platforms can make it more difficult to assign responsibility for moderation of harms, and potentially lead to arbitrage between the harms dealt with at a protocol level versus those addressed through user or community-based mechanisms.

How can we monitor and address the unintended consequences, biases, or side effects of safety features and ensure multiple perspectives, including those of marginalised groups, are considered when testing new features?

These technologies are often highly complex. How can users and regulators fill gaps in understanding so that transparency measures are meaningful?

**Possible new approaches or ways forward**

- Increasing dialogue between policy makers, civil society, and industry will be critical. Those making policies about new technologies and those creating them must be enabled to share information and concerns freely. The Call can help create trusted environments for those conversations to take place in a safe and open way.
- Members of our Community are already developing useful technical outputs which we can leverage, and promote, while injecting multistakeholder perspectives.
- The safety technology sector has been fast to adopt new technologies – particularly new AI tools. Bringing additional safety-tech companies and providers of open-source
tools into the Call Community could help to build relationships between safety providers and those who need assistance in countering TVEC.

- Other voices would also be useful to contribute to this conversation, including OSPs, experts in civil society, researchers with specialised knowledge, and young people (see section 4 below).
- The Community can provide a testbed and a place for dialogue on positive use-cases. These could include systems to more effectively detect TVEC for moderation purposes, use of models to identify gaps in existing moderation systems, support delivery of positive interventions, or contribute to media literacy by improving access to information.
- The Call Community can support risk identification and red-teaming activity. This will promote understanding of how new tools are implemented, where there may be unintended consequences for different communities, and how safeguards, transparency measures, and human rights impact measures can be effectively implemented when these tools are deployed.

2. Assisting smaller OSPs to eliminate TVEC online and respond to crisis incidents.

Why it’s important

In 2022, the Call Community observed there had been significant progress in dealing with the proliferation of TVEC and with incident response, and that it needed to continue developing tools and mechanisms to address adversarial behaviour on smaller OSPs that may lack the awareness, capacity and/or willingness to respond.

We also identified that, consistent with our commitment to an open internet, we needed to ensure support and tools were accessible to new entrants and smaller OSPs so that responsible approaches to TVEC did not become a barrier to competition and innovation.

We identified that crisis response mechanisms were often out of reach of many smaller OSPs, significantly limiting their ability to effectively respond to an online incident that might originate on or spill over to their services.

Recent progress and efforts by the Community

Canada’s funding for Tech Against Terrorism to expand the Terrorist Content Analytics Platform (TCAP) has led to a significant improvement in practical alerting capabilities for OSPs, leading to TVEC being removed faster and in more places online.

Support for smaller OSPs, including around regulatory requirements (e.g., the EC’s funding for Tech Against Terrorism and Violence Prevention Network to help with implementation of its Terrorist Content Online regulation, and the UK Home Office’s Capacity Building Programme).

A new tiered alerting framework for the TCAP, which includes crisis response assistance and hashing of URLs to address instances where links to externally hosted TVEC are shared on platforms.

An increase in GIFCT membership, with six new members joining in 2022-23.
Recent and upcoming free tools from Call-supporting OSPs to assist smaller providers with content moderation, such as Meta’s Hasher-Matcher-Actioner and Google Jigsaw’s new content moderation tool.

Increased dialogue with commercial safety technology vendors that are serving the needs of smaller OSPs through sophisticated classifiers and new technologies OSPs.

Continuation on the Free, Open, Secure Internet workstream - with the objective of bringing further technical community expertise into this work.

Difficult questions

A complex, expanding global regulatory framework is placing new obligations on all OSPs. How can we ensure that smaller providers are aware of their obligations and have the capacity to take effective action against TVEC, alongside other types of illegal and harmful content? How can the whole community better support smaller OSPs?

Smaller OSPs are not as easily able to ensure 24/7 incident response capability, transparency, and human rights protections and oversight in their trust and safety solutions. Are there practical ways the Call Community can help?

What is the most appropriate way of responding to a minority group of OSPs that chooses not to address TVEC on its platforms? How can we promote responsible action, without exacerbating or amplifying bad practice?

In some circumstances, and as a last resort, interventions at the infrastructure layer have been deployed to solve problems of TVEC. This has generally been a decision made by OSPs, consistent with their contractual terms of service. What could the Call Community contribute to a shared understanding of best practice and principles around infrastructure interventions, consistent with a free open secure internet, due process, and transparency?

Possible new approaches or ways forward

The Call Community is a leader in supporting smaller OSPs; continued funding and development of tools and resources is vital.

Consider ways to fill the gap in access to useful operational information (e.g., around threat analysis and the tactics of producers of TVEC).

Assistance for community-based open-source intelligence and research capability and creation of a trusted network of practitioners ready to assist with crisis incidents.

Work with smaller OSPs, Call partners, and the commercial safety technology sector to improve transparency and respect for human rights.

Improve 24/7 incident response support for smaller OSPs, building on the Terrorist Content Analytics Platform (TCAP) alerting infrastructure and closely integrated with existing protocols, including an updated Christchurch Call Crisis Response Protocol.

Open-Source human rights impact assessment toolkits for smaller OSPs to use.

Support and engagement from venture capital to promote human rights impact assessments as an essential part of trust and safety service offerings.
3. Deepening the evidence base on gender-based hatred & radicalisation to violence;

Why it’s important

Gender-based violence and gender-based online harassment and abuse (broadly gender-based harm / GBH) cause and are reinforced by individual and societal level harms on- and offline and can be a gateway to radicalisation to violent extremism.

There is a complex but correlated relationship between GBH offline and online - it does not exist within a vacuum. Rather, normalised sexism becomes magnified and amplified online. Acknowledging harm in this broader ecosystem of gender-based violence offline is important to understand and address this issue. GBH intersects with other forms of hate and processes of dehumanisation, such as racism and Islamophobia. In order to understand the drivers of GBH, an intersectional lens of gender is integral.

Misogyny and anti-LGBTQIA+ hate frequently operate as a common ideological connection across this spectrum of violence, and as a vector across different extremist ideologies. However, misogyny is also under-identified and under-prioritised as an ideological driver of TVEC in itself, reflecting an insufficient understanding of GBH, and violence towards women, girls and LGBTQIA+ people across societies globally.

Identifying how misogyny shapes ideological and corresponding algorithmic pathways to extremism enables earlier identification of, deterrence of, and interventions in radicalisation pathways, and better addresses GBH and gender-based online violence.

There is a strong policy interest from governments, civil society, and OSPs in how online harm translates to offline violence and vice versa including, for example in family violence, and harassment in professional and public environments.

There is a need to see accountability for victims.

Greater transparency and data access could make it easier to understand and address these questions.

Recent progress and efforts by the Community

The Call Community and the Institute for Strategic Dialogue (ISD) led workshops and events, including at RightsCon 2023, on assessing gender-based online violence (GBOV) risks found across design and functioning of platforms, and on OSP responses to these problems.

ISD has drafted a paper based on these discussions that identifies the links between misogyny, GBOV, and TVEC and makes recommendations for OSPs to mitigate GBOV and intervene in misogynistic pathways to radicalisation. This paper builds on recent research completed by civil society and academia and will complement other findings and recommendations that deepen the evidence base on the linkages between GBOV and violent extremism.

The EU Internet Forum’s handbook on borderline content includes some important analysis on GBOV, including examples drawn from Member States and OSPs that can be related to the violent extremism ecosystem.
The establishment of the Global Partnership for Action on Online Gender-Based Harassment and Abuse has helped to assemble expertise and institutional focus on these important issues, including supporting ISD’s paper.

Violence Prevention Network’s paper on Antifeminism provides insights into rhetorical strategies and connections to right-wing extremism.

Difficult questions

Misogyny and GBH are ideologically diffuse and manifest across a spectrum of actors, tactics, and harms, invoking varied community guidelines and policies. It is, however, linked to many categories of TVEC and research has canvassed how platform policies, approaches, and affordances exacerbate GBOV. What actions and mitigations can OSPs take to reduce GBOV on and across their platforms?

How can we differentiate between GBH that leads to radicalisation and that which does not? Where is the threshold for GBH in TVEC guidelines, given the broad spectrum of GBH? How can we address GBH and misogyny as a factor in radicalisation as part of viewing the broader ecosystem of GBH online and offline?

How do we acknowledge and disrupt the GBH and TVEC link through existing trust and safety tools, without compromising necessary data protection and rights to freedom of expression?

How can we achieve a higher level of standardisation and certainty when it comes to defining acceptable norms of behaviour online while maintaining a free, open, and secure internet?

To what extent are norms and policy settings for online platforms resilient to shifts in the extremism landscape (such as post-organisational violent extremism), and responsive to broader dehumanising narratives?

While misogyny has been identified as an integral component of many categories of TVEC, it continues to be under-identified by researchers and OSPs as a legitimate threat. How can we improve the identification of misogynist ideologies and GBH, especially when it is combined and embedded within other forms of online hate directed towards other marginalised groups?

How can the severity of GBOV be adequately assessed in content moderation and algorithmic pathways given the systemic under-prioritisation of GBH and of gender-based violence broadly?

Possible new approaches or ways forward

In TVEC policies, elevate the evidence that supports GBOV and misogyny as a driver for radicalisation to violence and identify clear impact goals for addressing GBOV.

Consider and operationalise existing recommendations developed by civil society (and especially victim-survivors of GBH) and academia to address GBH online and offline given research demonstrating its role in TVEC and radicalisation. Explore links with existing efforts to address dehumanising content, including CCAN’s 2022 output.
Work towards developing multistakeholder recommendations, deepen understanding, and test new methods for discouraging and preventing GBH online, including, for instance, work through the Christchurch Call Initiative on Algorithmic Outcomes.

4. Algorithmic outcomes, young people, & pathways to radicalization.

Why it’s important

The Call Commitments go beyond dealing with TVEC when it manifests, to address root causes, understanding how vulnerable individuals become radicalised, and how we can help prevent this from happening.

Many young offenders say they were radicalised online, suggesting there may be specific risk factors, and that interventions may need to be better tailored to the circumstances and interests of young people. This should include consideration of platform features other than recommendation/ranking systems.

There is a growing demand for data access and transparency, including through legislation.

Young people have important perspectives on new technology, its impacts on their wellbeing, and their autonomy and access to experiment and learn outside of their immediate environment. It is important that efforts to improve safety are carried out with consideration of their points of view, including their right to access and autonomy.

Recent progress and efforts by the Community

Call supporters have carried out significant research, including on how recommendation and ranking systems influence user sentiment and behaviour. This includes the EU Internet Forum and TrustLab’s report on algorithmic amplification, and the EU’s handbook on borderline content.

Many Call-supporting OSPs have shifted towards quality-based content ranking systems (i.e. away from engagement-based ranking). There have also been improvements to the automated and human systems that flag potential TVEC which reduce the risk of users encountering it online.

There have been further advancements in algorithmic transparency, for instance Meta’s publication of ‘system cards,’ which explain content recommendation systems and help users adjust their settings.

There is an increasing focus on ‘red teaming’ of online platforms and AI systems to evaluate for societal risks, including as a regulatory requirement.

Youth ‘councils’ and consultation bodies of different types are increasingly used to gather young people’s perspectives on a range of tech policy and online safety topics.

The Christchurch Call Secretariat is setting up a Youth Reference Group to provide advice to the Call Community. A youth engagement plan is in development, to ensure the safety of participants and enable them to make a meaningful contribution.
Difficult questions

People freely choose what they want to interact with online and, consequently, play a role in shaping machine learning outputs. There is no simple or single chain of causality, and so addressing the risk of radicalisation to violence on platforms requires nuanced understanding and responses, and concerted action. How can we build the right kind of trust to work in this nuanced way?

In 2022 Call Leaders emphasised providing platform users with tools that facilitate user choice and responsibility. How do we ensure these are available and that we understand risk/harm factors to help encourage positive choices?

How do we create meaningful dialogue with young people and how can we reflect their ideas / perspectives in our work?

How can we best address these nuanced questions in languages other than English and for cultures and communities that are not currently well-served by trust and safety efforts - including Indigenous, ethnic, disabled, or migrant and refugee communities?

How do we best ensure the public good and social responsibility are factored into our deployment of AI? Radicalisation of young people and GBH are important factors in that.

How do we equip users, particularly young people to be proactive about their own safety online?

Possible new approaches or ways forward

Identify clear impact goals for preventing violent extremism in specific areas relevant to young people, including e.g. the gaming sector.

Build expertise and community input into our work on the Christchurch Call Initiative on Algorithmic Outcomes (CCIAO – below)

5. Christchurch Call Initiative on Algorithmic Outcomes (CCIAO)

Why it’s important

The Call Community, policymakers, and academia share a strong desire to know more about how human/AI interactions on platforms affect behaviour, especially through content recommendations, and what role they may play in the radicalisation to violence of at-risk users.

As AI becomes more prevalent as a feature of online services, we need new ways of delivering meaningful transparency and public understanding.

There are limited options for third parties to research, compare and assess human/AI interactions on platforms. The process of doing research on user interactions can be expensive, complex, and difficult to replicate across time or between platforms.

Important user safety, privacy, and research ethics concerns need to be accounted for in any such work.
The CCIAO aims to develop new tools to empower research and help the community fulfill its objectives in this area. This includes building the evidence base around risks to young people, and impact measures for GBOV.

Recent progress and efforts by the Community

The US, New Zealand, Twitter, and Microsoft, working with OpenMined, DailyMotion and LinkedIn, have built and tested a system for remote researcher queries to be carried out securely and safely. This includes the rigorous testing of privacy tools by the data custodians and working through legal, ethical, and policy commitments.

The proof of function test has now been successfully carried out on multiple platforms, demonstrating that these tools could be used to support research about algorithmic outcomes.

Georgetown University’s Center for Security and Emerging Technology (CSET) has provided further support to this work, with a specific focus on on-chip transparency.

Key to the next stage will be working with the Call Community, building appropriate governance and ethics frameworks, adding more participating social media platforms, agreeing initial research priorities, and piloting initial research making use of this tool.

Difficult questions

How can we better connect cognitive and developmental experts into this work, e.g. to understand the apparent ubiquity of gender-based hatred and abuse in radical extremist groups?

There are new legal frameworks and voluntary measures promoting researcher access to data. Can the CCIAO make a positive contribution and, potentially, deliver better and more meaningful transparency outcomes as part of implementing those measures?

Possible new approaches or ways forward

Development of an ethics framework for researcher access, funding, and research priorities that reflect our commitment to human rights, privacy, and freedom of expression and association, to develop and enable use of new tools developed in CCIAO and better understand the progression of radicalisation to violence.

Expanding and scaling the CCIAO for use across the Call Community, involving members closely in direction and priority setting.

Engaging relevant experts to ensure research carried out in the next stage is rigorous and ethical, follows best practice, and contributes to our shared objectives.
6. Building, sustaining, and expanding the Call Community

Why it’s important

The ongoing commitment of the Call Community, and investment in the process and its outputs, is vital to fulfilling the Call Commitments.

Expanding and diversifying our Community can help further the mission of the Call and ensures the Call continues to deliver on its commitments as our environment continues to change.

Our work needs to be meaningful to affected communities and engage the expertise of people who understand these issues from diverse perspectives, including those of young people, women and girls, and LGBTQIA+ communities.

It is also vital that we continue to engage the technical community in our work, to ensure we live up to our commitment to preserving a free open, secure internet.

Transparency and trust are essential elements of an effective community; they enable us to share information and engage in good faith on shared problems and solutions.

Effective multistakeholder work requires sustained effort and commitment, and in turn helps us to make a positive impact. In a time of constrained resources, a collective approach can help us get more from our time and effort.

Recent progress and efforts by the Community

The Call Secretariat and Christchurch Call Advisory Network are onboarding new supporters, partners, and network members.

Over the last 12 months, we have hosted discussions on government transparency and reporting, and delivered guidance for our government supporters, building on the work of experts in the Community.

The Call also supported the Action Coalition on Meaningful Transparency to develop its Transparency Initiatives Portal, to connect members of the community to existing resources, reports, and upcoming events, as a basis for ongoing multistakeholder efforts to improve both technology company and government transparency.

New research outputs on the hybridisation of the threat landscape, e.g. ISD’s recent paper.

Call Community dialogue sessions on significant policy issues and engagement with the Community have been a useful means of addressing new challenges individual supporters might face.

We are assessing options for a Call Community platform to enable and build engagement among Community members.
Difficult questions

How can we expand while continuing to build strong Call Community relationships and a sense of trust and understanding between and across sectors?

How can we improve and professionalise the monitoring and evaluation of outcomes, and the fulfilment of Call Commitments by supporters?

What can the Advisory Network and the rest of the Community do to help encourage supporters in cases where they are acting contrary to their commitments?

How do we avoid or manage our way through conflicting measures that may make it more difficult for us to act globally, e.g. regulatory, or other developments? Can multistakeholder work help us reach better common understandings, and enable innovation in developing solutions?

How should the Call Community re-tool to confront the growing complexity and intersectionality of violent extremism, where other harm areas such as disinformation and child sexual abuse overlap with TVEC? How do we avoid scope creep, but also build a coherent approach to this evolving threat landscape?

How do we continue to centre in our work people and communities affected by TVEC?

There are growing attempts to discredit or politicise trust and safety practices. How do we maintain an inclusive conversation and maintain transparency and public trust in these efforts?

What are the best ways to meaningfully engage more diverse regions, cultures, and language communities in our work?

Possible new approaches or ways forward

Commission and deploy the Call Community platform to help promote better interaction, including information on accountability for delivery of commitments, and to promote and enable community-led efforts.

Promote dialogue on regulatory coherence, to better understand how we can support effective efforts to eliminate TVEC globally and avoid conflicting incentives or outcomes. This should include regular exchanges between different regions and interest groups, as well as better information sharing on new developments.

Reinforce the role of the technical community in the Call, to promote a free, open, and secure internet for good as a vital part of delivering our commitments.

Encourage adoption of the Call Community’s guidance on government transparency and reporting, alongside continued efforts towards improved industry transparency.

Work across the Call Community to develop community-wide benchmarks to assess, measure and evaluate our work to fulfil the Call Commitments and to build greater transparency, dialogue, and trust.
Build on the existing guidance by identifying and addressing known barriers to transparency and knowledge – in collaboration with the Action Coalition on Meaningful Transparency. This could include researcher access to data for the purposes of assessing and managing risk.

Create more resources to support smaller OSPs and civil society organisations to engage with the Call and deliver outputs meaningful to them.

7. Multistakeholder approaches to related issues

Why it’s important

There is an increasing effort by institutions, groups of governments or firms, and civil society groups, to define international norms and rules for technology. This is happening in the context of competing approaches to technology – ranging from an embrace of human rights and competitive, open global frameworks through to systems of top-down control and, at the extreme, authoritarian approaches to technology.

All sectors face questions about how they allocate resources to support different strands of effort internationally, and about the optimal models for making progress on complex, shared problems.

The Call takes a multistakeholder approach based on shared commitments and principles, executive-level leadership, and cooperative, community-led development of solutions and methods of implementation. This approach may be useful for addressing other, related digital technology challenges.

All stakeholder groups are stretched at present and are looking for ways to maximise their impact with the limited capacity they have.

Recent progress and efforts by the Community

Call Leaders asked for work to consider this issue, including the possibilities for the Call model to contribute in related areas.

France has drawn on the Call model to set up the Children Online Protection Laboratory to develop practical solutions to difficult problems. Working alongside existing multistakeholder frameworks including the Tech Coalition and WeProtect Global Alliance, it has embraced a collaborative, multistakeholder governance model for the initiative.

The Call Secretariat works to support linkages with other initiatives with aligned values and objectives, including Tech4Democracy, the Freedom Online Coalition, and the Global Partnership for Action on Gender-based Online Harassment and Abuse.

Some Call Community members have expressed a desire to reduce the transaction costs for stakeholders of engaging across multiple initiatives, and to support efforts on related or ‘Call-adjacent’ online safety issues in ways that draw upon the Call model and contribute to safety, freedom, and human rights online.
Difficult questions

What issues can/should be handled at a multilateral level? Will those be handled in a way that sufficiently includes civil society and private sector stakeholders?

How can multistakeholder initiatives like the Christchurch Call best relate to and contribute to multilateral efforts (including, e.g. the proposed Global Digital Compact)?

What approach should the Call Community take when such institutions or processes do not take an inclusive approach, or appear to have a lower level of ambition and efficacy than multistakeholder efforts?

What should be the role of multistakeholder approaches, like that of the Call, in helping address issues with the responsible deployment of technology in a way that brings together governments, the technical community, civil society, and industry? Are there suitable opportunities in specific areas? How do we best ensure these are effective, efficient multistakeholder processes?

How do multistakeholder perspectives contribute to improving regulatory work, including as codes, standards, and benchmarks are developed in different jurisdictions?

Can we bring greater coherence and accessibility to these topics?

Possible new approaches or ways forward

New institutions or structures to help coordinate multistakeholder efforts, share information, and make it easier for small institutions, governments, and organisations to participate in shaping technology norms.

Information sharing for situational awareness across the Call Community on processes we are involved in. Supporting advocacy for multistakeholder approaches in diverse processes and issues and coordinating multistakeholder input where needed.

Other issues raised

As part of the Town Hall meetings, and through written input, Community members raised some additional issues to be considered. These included:

- Dehumanisation as an overarching theme of harmful content / conduct.
- A continued focus on testing, updating, and implementing the shared crisis response protocol, and integrating transparency, oversight, and due process protections across all the protocols active in the Community; interoperability of crisis response mechanisms; and
- The imperative of communicating the outcomes and achievements of the Call.
Conclusions

The supporters of the Christchurch Call have committed to fulfil the vision of eliminating terrorist and violent extremist content online, consistent with human rights and a free, open, secure internet as a force for good. The Call is supported by a strong Community that includes civil society and partner organisations working alongside the supporters and the Secretariat to achieve these goals.

Continuing to build a strong, cohesive, diverse, and effective Call Community remains the most important enabler across all the Call work. The trust and commitment that resides in our Community is our most precious resource. Delivering and being accountable to our Commitments and engaging the expertise from across our community helps to build trust and commitment. We have outlined actions on which there is broad agreement about how to proceed, including adding new voices to our work, developing monitoring benchmarks, and making additional efforts to build meaningful transparency and dialogue.

There is a clear priority attached to improving our incident response systems, lowering the barriers to participation, and improving the capabilities available for smaller OSPs to participate in the task of eliminating TVEC. In the context of a growing safety technology ecosystem, the Call Community should help to promote human rights standards and ensure a sufficient mix of commercial, open-source, and pooled capability is available, and that this is effective, up to date relative to the technical challenges, and enables all actors to play a constructive role should they wish to. This work will need to engage the technical community and to uphold our commitment to a free, open, and secure internet. It will need to adapt as new technologies become widely available and the threat landscape continues to hybridise and evolve.

Looking at this year’s Summit, and its implications for our work in 2024, the Call community has identified important challenges that require discussion:

1. There is a strong consensus from across the Call Community that we should devote time to considering how new technology and, particularly, new foundation AI models affect the effort to eliminate TVEC online. Working within the Call Community, these issues can be addressed with specificity in a trusted environment that includes industry, government, and civil society expertise. These efforts will need to be carried out with reference to other international efforts to develop commitments and safety measures for AI more generally.

2. Call supporters are committed to understanding algorithmic outcomes, including platform features and interactions that can lead to real world harm, radicalisation to violence or exposure to TVEC. A discussion of these issues requires an understanding of how user behaviour and offline inputs factor in, and it requires consideration of other contributing elements, such as GBOV (including anti-LGBTQIA+), as drivers of radicalisation to violence. New transparency requirements and tools can assist in deepening understanding of these issues. To fulfil Leaders’ objective of developing a new generation of community-driven online interventions, the Call Community will need to agree on ways to support and deliver targeted interdisciplinary research, enriched dialogue with young people, and deeper technical collaboration. This will require additional effort and resources, and a bigger coalition to lean into the work.

3. Civil society, governments, and industry are all struggling with the crowded international agenda. There is no clear alignment about where to go to develop norms and rules for the responsible deployment of technology. It is desirable to promote multistakeholder engagement on these topics and to support initiatives that bring stakeholder groups together and operate...
based on a common commitment to human rights law. Where existing initiatives, such as the Christchurch Call, are playing a useful role, it may be desirable to consider ways they can be made resilient to change, and whether they might usefully support multistakeholder work on related issues. The Call Community would like to consider ways to increase mutual support of, and coherence across, related initiatives to provide greater visibility and the ability to coordinate our efforts for impact, and may wish to consider a small number of related issues on which to focus next.
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

<table>
<thead>
<tr>
<th><strong>Large Language Models (LLMs)</strong></th>
<th>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’.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Generative AI</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Deepfakes</strong></td>
<td>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.</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decentralized Web (Dweb)</td>
<td>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.</td>
</tr>
<tr>
<td>Decentralized social network</td>
<td>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.</td>
</tr>
<tr>
<td>Blockchain</td>
<td>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).</td>
</tr>
<tr>
<td>Peer-to-peer Connectivity</td>
<td>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.</td>
</tr>
<tr>
<td>Decentralized Autonomous Organizations (DAOs)</td>
<td>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.</td>
</tr>
</tbody>
</table>

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 use 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

<table>
<thead>
<tr>
<th>Commitment</th>
<th>Implications of new technologies</th>
</tr>
</thead>
</table>
| 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**<br>Potential for generative AI and immersive tech harms to contribute to drivers of TVE including social division and deteriorating trust in information.<br>Potential for generative AI and immersive tech to train and provide knowhow to TVE actors.  
**Opportunities**<br>Uses of these technologies to create and share effective counter narratives, and for education/media literacy purposes.<br>This may include technical solutions such as flagging or providing context for artificially generated content (e.g. Content provenance technologies, context notes on Twitter...).<br>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**<br>TVE actors may use new technologies to create new forms of TVEC and spread these more widely and efficiently.  
**Opportunities**<br>Work across Call Community to develop effective and rights-respecting ways of identifying and removing this content – for example through AI-enabled detection.<br>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**<br>New forms of content and ways of distributing it may make it more difficult to stop the spread of TVEC.<br>Risk of over-correcting and harming FOE.<br>AI-enabled content moderation may lack transparency.  
**Opportunities**<br>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. |
### 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.

### 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.

### 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.

<table>
<thead>
<tr>
<th><strong>Risks</strong></th>
<th><strong>Opportunities:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Generative AI accelerates user journeys towards TVEC due to increased personalisation.</td>
<td>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.</td>
</tr>
<tr>
<td>Proliferation of content and networks, including via DWeb, may make it difficult to implement interventions.</td>
<td>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.</td>
</tr>
<tr>
<td>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.</td>
<td>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.</td>
</tr>
</tbody>
</table>
Community Guidance: Government Transparency and Reporting
Table of contents

Introduction ........................................................................................................................................ 0
About this guidance ............................................................................................................................ 1
Scope .................................................................................................................................................. 2
Why is government transparency important? ..................................................................................... 3
Key Pieces of Community Advice .................................................................................................... 5
  CDT’s framework for meaningful transparency .............................................................................. 5
  Santa Clara 2.0 .................................................................................................................................. 5
  Tech Against Terrorism’s Guidelines for Government .................................................................. 6
Government Transparency Reporting .................................................................................................. 6
  Principles .......................................................................................................................................... 6
  Categories and Content ................................................................................................................... 8
Examples and Good Practices ............................................................................................................. 10
  1: The Internet Hotline Centre, Japan ........................................................................................... 10
  2: Independent Review, ARCOM, France ...................................................................................... 10
  3: EU TCO Framework for Transparency ...................................................................................... 11
  4: Transparency Reporting: EU Internet Referral Unit, Europol .................................................. 13
  5: Transparency Reporting: Department of Internal Affairs, New Zealand .............................. 13
  6: Transparency Reporting, Ofcom, UK ......................................................................................... 14
Introduction

It is a priority for the Christchurch Call Community that supporters – both governments and tech companies – provide meaningful transparency around how they are acting to eliminate terrorist and violent extremist content online while respecting and promoting human rights and a free, open and secure internet. This objective was clearly articulated and set in the Community work plan for transparency endorsed by Leaders in 2021.

To date, most efforts to improve transparency have focused on and been from within tech companies. This started in the early 2000s in response to growing civil society concerns about the relationships between these companies and governments.1 The Berkman Klein Center for Internet and Society developed the earliest version of the Lumen database more than 20 years ago, in 2002. In 2010 Google published the first report on government requests it received. Interest in tech companies’ disclosure of this sort of information only intensified through the mid-2010s, following the Snowden revelations.

Transparency reporting is now a well-established and reasonably widespread practice among the tech companies that support the Call. This reporting has developed beyond basic data on government legal requests to cover company content policy and enforcement, and information operations disclosures.2 There is also expert guidance from academia and civil society3 and from various regulatory and voluntary initiatives, including some specific to terrorist and violent extremist content, that aim to make tech companies’ transparency reporting more widespread, consistent, and meaningful.4

By contrast, despite transparency being a well-established democratic principle, there has been less emphasis on government transparency in relation to content moderation.5 Only a handful of the Call’s government supporters regularly publish transparency reports.6

Yet government transparency is a subject of legitimate interest and concern to all Community members. It is not possible to understand the whole online content ecosystem without visibility of governments’ influence in determining what stays online and can be easily accessed, and what gets removed or restricted, whether because it is unlawful or a breach of a company’s terms of service.

1 For more, see History of Transparency Reports – Trust & Safety Professional Association (tspa.org)
2 Integrity Institute Shining a Light on Platform Transparency Best Practices July 2023
4 Examples include Tech Against Terrorism’s mentoring to meet GIFCT membership criteria, the OECD voluntary transparency reporting framework and transparency reporting provisions in the EU Terrorist Content Online Regulation.
5 See e.g. paragraphs 50-51 of the Office of the Privacy Commissioner’s 2017 report, Transparency Reporting, which notes the difficulties of getting comprehensive reporting out of government agencies.
6 Noted in Vogus, C and Llanso, E “Making Transparency Meaningful: A Framework for Policy Makers”, Centre for Democracy and Technology, December 2021. This small group includes Europol’s Internet Referral Unit and the Department of Internal Affairs in New Zealand. Australia’s e-Safety Commissioner’s annual report also includes some relevant metrics.
Government involvement in content moderation can give rise to significant human rights risks. Those risks include the potential for circumventing the law and due process, introducing bias into company policy- and decision-making, using tech companies as proxies to silence legitimate grievances and dissent, and interfering with users’ access to effective remedies. Disclosing information about how, why, and when governments exercise influence over digital content can help mitigate these risks by building trust, fostering accountability, and facilitating access to remedy for affected users.

All Call supporters have committed to work together with civil society to improve transparency and ensure that Call commitments are implemented in a manner consistent with a free, open, and secure internet, and international human rights law.

Call-supporting governments have also made relevant commitments elsewhere. For example, all have signed up to Sustainable Development Goal 16. Most are members of the Open Government Partnership and have therefore committed to increase the availability and usefulness of information about government activities at all levels. Many are members of the Freedom Online Coalition, which has called on governments everywhere to promote transparency and independent, effective oversight around the use of content take-down notices, restrictions on online content or user access, and similar measures, while committing to do the same and to preserve and strengthen the multistakeholder model of internet governance.

The focus of this guidance is on government transparency, and reporting in particular. That is not to suggest that there is no need for further improvement on the tech company side. The Integrity Institute judges that “to date, voluntary transparency efforts from the industry have not provided a satisfactory level of transparency”. As key pieces of regulation come into effect, the Call Community needs to evaluate their impact on transparency and continue to lift standards and address gaps, including around the impacts of algorithms on user journeys, and safety policies and tools for emerging technologies.

About this guidance

In 2023, the Call Community started a series of discussions on government transparency. The focus was on things governments could do to improve transparency in their efforts to detect terrorist and violent extremist content online, and associated legal requests and other referrals to tech companies to remove, restrict access, or otherwise take action on it and on associated accounts.

The objectives of the Community sessions were:

- To identify the key stakeholders for government transparency, what they need to know, and how best to deliver that information in a timely and accessible way;

---

7 Office of the UN High Commissioner for Human Rights “Moderating online content: fighting harm or silencing dissent?” July 2021
8 Pielmeier, J and Sheehy, C “Understanding the Human Rights Risks Associated with Internet Referral Units,” Feb 2019
9 See Oversight Board overturns Meta's decision in "UK drill music" case | Oversight Board
10 SDG 16 is “Promote peaceful and inclusive societies for sustainable development, provide access to justice for all, and build effective, accountable and inclusive institutions at all levels”.
11 FOC-recommendations-consensus.pdf (freedomonlinecoalition.com)
12 Christchurch Call Initiative on Algorithmic Outcomes | Christchurch Call
To socialise existing guidance and identify any gaps;
To explore some current approaches to government transparency;
To develop Community guidance to support more consistent and meaningful transparency by governments, with good practice examples.

Meaningful transparency is most likely to emerge in a framework or regime that encompasses different actors, types of information and data, modalities, and formats. Governments have a role in shaping the overall framework, through voluntary and cooperative initiatives like the Christchurch Call, through co-regulation and regulation. Governments, as actors in content moderation, also need to provide transparency around their own policies, processes and activities within the framework. That transparency can take various forms, depending on the audience and purpose.

This guidance begins by considering an overall framework for transparency, and the different forms that it can usefully include, before honing in on one particular form: reporting. It synthesises some principles for government transparency reporting as well as key categories and content to cover, as put forward in the nascent body of expert advice published by Community members and articulated in our Community sessions. The guidance also contains case studies that highlight some existing good practices among the Call’s government supporters.

Scope

The key relevant Call commitments are set out in the table below.

Table 1: Call Commitments Relevant to Content Moderation and Transparency Reporting

<table>
<thead>
<tr>
<th>Online service providers commit to:</th>
<th>Governments commit to:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set transparent terms of service, describing policies on terrorist and violent extremist content and consequences for users who share it.</td>
<td>Ensure the effective enforcement of applicable laws that prohibit the production and dissemination of terrorist content.</td>
</tr>
<tr>
<td>Take transparent, specific measures to prevent upload, detect and remove such content.</td>
<td></td>
</tr>
<tr>
<td>Implement regular and transparent reporting on the quantity and nature of such content being detected and removed.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Online service providers and governments commit to:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognise the important role of civil society in supporting the Call, including working with supporters to increase transparency and to implement the commitments in a manner consistent with international human rights law and a free, open and secure internet.</td>
</tr>
</tbody>
</table>

Transparency, participation, and accountability in the policy-making and legislative processes that lead to the passing and enactment of the “applicable laws” is particularly important.\(^{13}\) While this is an essential part of open government, how to do that well is a huge topic on which much has been published and it is beyond the scope of this guidance. Rather, the focus here is on transparency around implementation and enforcement of existing laws and policies focused on the availability of

\(^{13}\) Tech Against Terrorism calls for greater transparency in legislative processes in the Online Regulation Series 3.0 (2023 edition).
TVEC on information and communication technology services. Also in scope are government requests and referrals under company terms of service.

In some countries, it is a criminal offence for an individual to possess, produce or distribute illegal terrorist content. Governments and online service providers that support the Call commit to ensure appropriate cooperation with law enforcement agencies for the purposes of investigating and prosecuting illegal online activity regarding detected and/or removed terrorist and violent extremist content. Where the agency that is responsible for issuing a takedown notice to an online service provider hosting illegal terrorist content also has the power to investigate the individual who posted it, transparency (e.g. reporting) would usefully cover requests for user data. In countries where relevant investigative powers sit with other agencies, it would be useful to include or at least reference relevant data or reporting by those agencies. However, government demands and requests to tech companies for other law enforcement and security purposes are beyond the scope of the Call and this guidance.

**Why is government transparency important?**

Transparency is fundamental to building citizens’ trust and participation in democracies, and to effective and accountable government, consistent with rule of law and the protection and promotion of all human rights.

Democratic states typically have well-developed institutions and processes to create transparency about executive government, as a prerequisite for citizens to hold elected and non-elected officials accountable. Those include the hearing and inquiry powers of Legislatures and Parliament, independent oversight bodies like auditors and Ombudsmen, departmental or agency annual reporting requirements, and freedom of information and privacy laws. However, the information provided, e.g. in government departments’ annual reports, is typically high level. It may lack the specificity and granularity to enable groups and individuals to understand, assess, debate and challenge laws and activities that interest them or impact their rights.

The Christchurch Call aims to eliminate terrorist and violent extremist content online while protecting and respecting human rights and a free, open, and secure internet. It is well recognised that government laws, policies, programmes, and actions in this space can impact – directly and indirectly, positively and negatively – on the full enjoyment of all fundamental rights and freedoms.

In the Call Community, we seek to centre the voices and experiences of victims, survivors, their families, friends, and communities. The rights to life and to security of the person must be protected and promoted through effective action against terrorist and violent extremist content online. This is a fundamental role for governments. However, individuals and communities most impacted by terrorism and violent extremism are also often negatively impacted by counter terrorism activities. BSR has noted, e.g. that the victims of overbroad restrictions on online speech have tended to be from Muslim communities, which in turn impacts other rights and freedoms, including access to information and to effective remedy. It is also the case that many states abuse counter-terrorism laws to silence and oppress opponents and minorities.

If transparency is important for individuals and groups unfairly targeted and stigmatised by counter-terrorism, it is also important for those who feel unfairly targeted. Post-pandemic, many countries have experienced increasing distrust in government and ‘big tech’. Sometimes, there is legitimate concern and debate about authorities’ relationships with digital platforms, and the possible flow-on

---

14 See e.g. BSR “Human Rights Assessment: Global Forum to Counter Terrorism”, pg 27 July 2021
effects on freedom of speech. In other cases, mistrust and anti-authority views can be exacerbated by the dis- and misinformation and conspiracy theories that can flourish in the absence of open, accessible, and specific information from governments.

This is why, at a high level, it is important for governments to provide transparency around their laws, policies and actions to remove or restrict terrorist and violent extremist content online.

More specifically, such transparency:

- Informs the public, including vulnerable communities and individuals, about what the government is doing to address terrorist and violent extremist content online and prevent harm.
- Deters officials from acting outside the law or applying it unfairly, to censor speech, to block access to information, or to obtain user data, illegally, for illegitimate reasons or in overly broad or discriminatory ways.
- Builds public trust and confidence that laws are being implemented and enforced fairly and effectively.
- Empowers everyone to exercise their rights online and – in conjunction with legislative and independent oversight bodies, the judiciary, media, and civil society – to assess and hold governments accountable for wrongful decisions, illegal actions, and ineffective policies and programmes.
- Leads to better laws, policies and practices over time, by forcing internal reflection, establishing an evidence base on how well the status quo is working, and generating improvements informed by diverse perspectives.
- Enhances domestic legitimacy while also demonstrating adherence to international human rights law and counter-terrorism standards for a global audience that includes other states.

Clearly, government action to combat terrorist and violent extremist content online impacts online service providers and users all around the world. In the context of the Christchurch Call, which is grounded in a global multi-stakeholder approach, it is incumbent on government supporters to improve transparency alongside the tech sector, especially as they seek to impose mandatory reporting and other transparency obligations on it. Such transparency is the foundation for genuine dialogue in our Community, for the development of mutual understanding and trust, and for effective collective action.

---

15 See e.g. Tech Policy Press “What Should We Know About Government Influence on Content Moderation” February 2023; also
Key Pieces of Community Advice

Over the last couple of years, experts in industry, academia and civil society have begun to publish guidance specifically for governments, focusing on their role in enabling and creating meaningful transparency. In this section, we summarise these existing resources.

CDT’s framework for meaningful transparency

In 2021, the Centre for Democracy and Technology (CDT) put forward a framework for policymakers to help make transparency more meaningful for stakeholders.\footnote{Vogus, C and Llanso, E “Making Transparency Meaningful: A Framework for Policy Makers” Centre for Democracy and Technology December 2021} It recognises that there are many different forms of transparency, produced by different actors, for different audiences and purposes. Those forms include:

- Transparency reporting containing aggregate data and qualitative information about moderation actions, disclosures and other actions concerning user generated content and accounts.
- User notifications about government demands for moderation of their content or their data.
- Access to data held by intermediaries for independent researchers, public policy advocates and journalists.
- Public facing analysis, assessments, and audits of tech companies with respect to user speech and privacy.

Each of these forms of transparency has its own strengths and limitations. Brought together in a considered way, these approaches can reveal the ‘big picture’ i.e. how the policies, processes, decisions and actions of governments and tech companies interact, shape online content, and affect users.\footnote{ibid}

As part of the Call’s work on government transparency, CDT suggested the Call framework include descriptions of laws, policies, and administrative arrangements, which could be provided on a government department’s website or as part of periodic reporting. CDT also suggested that governments could directly and proactively provide all these forms of transparency, except user notifications. Although we are not aware of any that provide researcher access to data in practice, there are certainly relevant precedents\footnote{See e.g. New Zealand Police Research – External Researcher Requests} and this is an area that could be explored by the Community in more detail. Governments also shape the overall framework (including mandating or blocking tech companies’ provision of user notifications), through regulation, co-regulation and in voluntary cooperative initiatives with the tech sector - see EU example below.

Santa Clara 2.0

The Santa Clara Principles on Transparency and Accountability were first developed by a group of human rights organisations and academics – including CCAN members – in 2018, focusing on tech companies and the rights of their users, and the information needs of academics and civil society in helping defend and promote those rights. The principles were updated in 2021, with one aim being to address the role that governments play alongside tech companies in content moderation and the risks and concerns that arise.
Santa Clara 2.0 says that there should be transparency around government involvement in the development and enforcement of a tech company’s rules and policies. Users should know when a state has been involved in action on their content or account, whether on legal grounds or under the company’s own terms of service. This is a foundational principle that applies to all companies, no matter the type, size or age. Tech companies need to provide user notifications to enable this kind of transparency.

The updated principles also state that governments should recognise and minimise their roles in obstructing tech company transparency and must also provide transparency around their own involvement in content moderation decisions, including demands for content removal or restriction. Transparency reporting by governments is therefore also an important part of the overall framework for delivering against the Santa Clara principles.

Tech Against Terrorism’s Guidelines for Government

In 2021 Tech Against Terrorism published guidelines for government transparency reporting on counter-terrorism efforts. They complement separate guidelines for tech company transparency reporting and are based on a ‘waterfall’ model that covers:

- **Legal basis** i.e. the law(s) under which government agencies refer content or accounts to tech companies for action, and/or request user data.
- **Processes and systems** that the government agencies use to detect terrorist and violent extremist content and accounts, and to refer them for action either on the basis of local law or the tech company’s terms of service. This also covers record keeping and redress mechanisms.
- **Outcomes** as revealed by quantitative metrics for referrals and requests; broken down by means of discovery, ground for action, company and outcome; and appeals and complaints received as well as the outcomes.

The guidelines contain templates and examples of the qualitative and quantitative information that should be included in a government transparency report. They go a long way towards operationalising the Santa Clara principles as they relate to terrorist and violent extremist content. The guidelines are therefore a particularly useful resource for Call supporter governments and the wider Community.

**Government Transparency Reporting**

**Principles**

Drawing on the above resources, the wider literature and our Community sessions, the following principles are intended to guide government supporters of the Call in their efforts to promote transparency in the form of reporting:

- Consider the whole online content ecosystem, the different actors and stakeholders involved, and how best to provide them with relevant information and data. Think about which elements of CDT’s framework for meaningful transparency are most relevant to promoting understanding, accountability, and continuous improvement in your country, and in the multistakeholder forums your country participates in. Consider how periodic reporting fits in the framework and adds value for stakeholders.
Identify key audiences, internal and external, domestic and international, and consult them on the purpose and their specific use(s) of information and data that will go into your reporting, as the ‘why’ should inform the approach and content.\(^{19}\)

Contextualise the information and data in the report. Describe the different ways and places your government interacts with tech companies to address TVEC online. Take stock of and explain the unique laws and administrative arrangements for government requests and referrals, and transparency, in your country.

If multiple departments or agencies make orders and requests, or otherwise engage with companies on content moderation, aim to ensure appropriate coverage of these agencies’ efforts, depending on the information needs of key stakeholders. Reference relevant information and reporting from these other sources, as well as from any independent review or oversight bodies. Note that the Santa Clara Principles call for consolidated reporting across different kinds of illegal or policy-violating content (e.g. adult, copyright, defamation, hate, privacy, violence). Consolidated reporting of this kind will be particularly valuable for some stakeholders. Others will derive greater value from subject-specific reporting, which may be more detailed and regular.

Acknowledge and seek to address limitations and barriers to comprehensive reporting, including legal, financial, and practical impediments.

Look at tech companies’ transparency reports and the data they provide on your government’s orders and requests. Seek to understand and explain differences in scope between company and government reporting and apparent anomalies.

Make information and data accessible and easy to understand. Use plain language as much as possible. Define key terms, especially technical ones. Analyse and interpret the data. Describe the nature of content being referred. Explain how any incidents or trends during the reporting period impacted on the nature and volume of content referred to tech companies. Avoid spin.

Provide regular updates e.g. on an annual or six-monthly basis.

Strive for consistency across reporting periods. But do not let that prevent innovations in approach that will better serve stakeholders and their interests. Make sure to highlight any changes in the context for or approach to reporting since the last one and explain any impacts on the information presented.

Make a start and don’t let the perfect be the enemy of the good. As is the case for tech companies, producing a government transparency report can be resource intensive and require new systems and tools as well as coordination across multiple units within and beyond a specific department.\(^{20}\) Governments will be starting from different places.

Look to continuously improve. Actively seek feedback from stakeholders. Look to the leading edge of tech company practice, too, for approaches that may better serve them. Tech companies are e.g. evolving away from static PDF reports to comprehensive transparency centres, with a range of interactive, accessible, and machine-readable methods of disclosing information and data.\(^{21}\)

---

\(^{19}\) Action Coalition on Meaningful Transparency [https://www.meaningfultransparency.tech/post/reporting-summary](https://www.meaningfultransparency.tech/post/reporting-summary)

\(^{20}\) ACT Discussion Summary: What goes into preparing a tech transparency report? [meaningfultransparency.tech](meaningfultransparency.tech)

\(^{21}\) ibid
Recognise the inherent limitations of transparency reporting\textsuperscript{22} and convene multi-stakeholder discussions to enable trusted information-sharing with key stakeholders, as well as dialogue around your government’s laws and activities, and the impacts on safety, security, and other fundamental rights and freedoms.

**Categories and Content**

The table below provides a possible structure and overview of the categories and types of content that governments may consider providing e.g. in transparency reports. It is based on the recommendations in Tech Against Terrorism’s guidelines and advice published by other experts in the Community and beyond\textsuperscript{23}, tailored to the Call context and scope of this guidance, and reflecting the discussions in our Community sessions. It is recommended that governments refer to the source documents for detail and nuance as they work to develop a transparency framework and reporting.

![Table content](image)

\textsuperscript{22} Integrity Institute *Shining a Light on Platform Transparency Best Practices* July 2023: page 15.

against terms of service?

Describe how your laws and policies are operationalised and enforced in practice

- What approaches, systems and tools do the relevant officials and/or state agents use to discover terrorist and violent extremist content online?
- How do they submit orders and requests to tech companies?
- How do they make referrals against terms of service?
- Are there any internal review mechanisms to ensure orders, requests and referrals are lawful and well-founded before they are sent?
- What data is recorded and stored about discovery of online terrorist and violent extremist content online, and in relation to orders and requests, and referrals, sent to tech companies?

Describe any structures, mechanisms and provisions for independent oversight, review and redress

- What structures or processes exist to provide independent oversight of the department(s) and/or agencies involved? Are orders, requests and referrals externally reviewed? If so, how?
- What channels and forms of redress are available to users whose content or accounts or data has been wrongfully subject to an order, request or referral?
- How do you support companies in providing transparency and facilitating redress in cases where their content or account has been wrongfully actioned or data wrongfully obtained?

Provide data and explanations on activity and outcomes

- How much terrorist and violent extremist content has been proactively discovered using the various processes and systems? How much has been discovered because of reports from the public?
- Quantity of content that violates law or policy, disaggregated by: source of discovery; nature of violation; terrorist group and designation status, or ideological category; and company.
- Number of legal content-related orders and requests made to tech companies, disaggregated by: company; nature of legal violation; terrorist group and designation status, or ideological category; action ordered or requested; outcome.
- Number of referrals and requests made outside legal channels, disaggregated by company; nature of policy violation; terrorist group and designation status, or ideological category; action requested; outcome.
- If relevant, number of prosecutions against tech companies and/or individuals, disaggregated by: nature of legal violation; outcome, including penalty.
- Appeals or other challenges made against government or state agent orders or requests, disaggregated by: company; type of order or request (legal or ToS referral); success rates; remedies.

Other useful information

Consider including information on:
Examples and Good Practices

1: The Internet Hotline Centre, Japan

In Japan, The Internet Hotline Centre (IHC) is an independent non-governmental organisation responsible for assessing potentially illegal or harmful content online and referring it to online service providers for action.

Any internet user or member of the public can flag such content to IHC. It then assesses the content against operational guidelines and refers illegal or harmful content to the relevant online service provider(s) or website administrators for action. The IHC may request that the online service provider removes or blocks the content if it is illegal or takes voluntary action on harmful content based on terms of service. Although there is no specific prohibition on terrorist or violent extremist content, it would fall under categories of harmful content related to depictions of / incitement to serious violent crimes, and/or sale or manufacture of weapons.

IHC also reports content to the National Police Agency, where it appears to violate national law or be related to a significant crime or represent an imminent threat to life (e.g. suicide).

IHC publishes six-monthly and annual data on its website, including the total number of reports it has received, broken down by the type of illegal or harmful information and action taken by IHC (including reports to Police) and online service providers.

The IHC’s work is periodically reviewed and evaluated by an Operating Committee of five external people from academia and civil society.

Good practices

- State actors are not directly involved in content moderation.
- IHC publishes its operational guidelines (including definitions of different kinds of illegal and harmful content) and regular (six-monthly) on activities and outcomes.
- There is provision for independent review of the IHC’s work, and multistakeholder evaluation of subtle decisions that can feed into updates to the operational guidelines.

2: Independent Review, ARCOM, France

ARCOM is the French Regulatory Authority for Audio-visual and Digital Communication. Its role is to guarantee freedom of expression and access to information online.

Since 2014, the Central Cybercrime Prevention Office (OCLCTIC) of the Judicial Police has had the power under French law to issue take down notices to online service providers hosting terrorist
content. If the online service provider does not remove the content within 24 hours, OCLCTIC can ask that the content be blocked or dereferenced.

To enable fast action there is no need for OCLCTIC to obtain a court order for these requests. Instead, the Board of ARCOM is empowered to appoint a qualified member of its board to oversee OCLCTIC operations and review all its requests to ensure compliance with the law.

The current appointee is Laurence Pécaut-Rivolier, a judge and member of the Court of Cassation. She may recommend that OCLCTIC revoke a notice that is not well-founded in law and ask that the relevant content be reinstated. Where there is a difference of opinion between her and OCLCTIC, she may refer the matter for judicial review.

In this role, Ms Pécaut-Rivolier must report annually to the Government and Parliament on the number of orders issued by OCLTIC and the number of pieces of terrorist content taken down or blocked or dereferenced as a result. It is her practice to include information about the instances where she has recommended revocation of a notice and how they concluded.

The 2022 ARCOM report is published online in full and in summary form. Ms Pécaut-Rivolier also convened a meeting of industry and civil society stakeholders to discuss the report and its conclusions.

Under the EU TCO Regulation, Ms Pécault-Revolier is the competent authority for scrutinising cross-border removal orders against French hosting service providers. Arcom oversees the implementation of specific measures by HSPs exposed to terrorist content, and imposes any penalties. ARCOM will issue transparency reports under Article 8, and has contributed data and information to France’s national reporting under Article 21.

**Good Practices**

- Transparency reporting on ARCOM’s independent oversight and review functions means the French Parliament and public can be confident that OCLCTIC powers are being exercised in accordance with the law and fundamental rights and freedoms.
- Reporting includes information on:
  - Legislative framework and administrative arrangements for removal orders, oversight, review and redress.
  - Activities and outcomes, e.g. characteristics of terrorist content subject to removal orders in a given year.
  - There is an opportunity for multistakeholder dialogue around OCLCTIC and ARCOM activities and outcomes each year.

**3: EU TCO Framework for Transparency**

The regulation addressing the dissemination of terrorist content online came into application in June 2022. It empowers designated competent authorities in EU Member States to send removal orders to hosting service providers requesting they remove or disable access to terrorist content. They have one hour to comply. The Regulation also empowers Member States to decide, based on objective factors, whether a hosting service provider is exposed to terrorist content and therefore must take specific measures to address it.

The Regulation establishes a framework for transparency around the use of these powers and actions taken by hosting service providers, as an integral part of the safeguards for human rights. Within that
framework, there are transparency and reporting obligations for hosting service providers, competent authorities, Member States and the European Commission.

The European Commission must publish an online register of all designated competent authorities, based on information from Member States.

Under Article 7, hosting service providers must set out clearly in their terms of service their policy for addressing the dissemination of terrorist content and how they implement it. Any hosting service provider that has received a removal order and/or been required to take specific measures to address terrorist content on their service must issue a transparency report containing the number of removal orders received, details of specific measures taken, complaints, judicial or administrative reviews, and outcomes. They are also asked to provide a meaningful explanation of the functioning of specific measures.

Under Article 8, national competent authorities must publish annual transparency reports mirroring those required from hosting service providers and containing information on the number of removal orders sent, decisions to require specific measures against terrorist content by hosting service providers, judicial or administrative reviews, outcomes, and any penalties imposed.

Under Article 11, hosting service providers must notify users whose content has been taken down or blocked because of a removal order. The content provider can request reasons or a copy of the order. A competent authority can delay such notification for six weeks, where necessary and proportionate, to enable investigation and prosecution of a terrorist offence. That can be extended for another six weeks.

Under Article 21, Member States must bring together information from both hosting service providers and their national competent authorities and submit an annual monitoring report to the European Commission on the implementation of the regulation in their territory.

Under Articles 21 and 22, the European Commission must work with Member States to establish a programme for monitoring the outputs, results and impacts of the regulation and submit an overall implementation report to the European Parliament and the European Council in 2023.


The transparency reports from hosting service providers and competent authorities, and the European Commission’s implementation and evaluation report, are public. The Regulation sets out the minimum required information, although hosting service providers and national competent authorities can provide more.

Europol has developed a digital platform, PERCI, to connect competent authorities and hosting service providers and help them fulfil their obligations under the regulation. Competent authorities issue removal orders and referrals through the platform, enabling coordination across Member States and with Europol. Hosting service providers enter information about their response e.g. the time at which the content is taken down or blocked. This will support the production of annual transparency reports. PERCI will be further developed to provide meaningful information about actions taken by HSPs because of a referral or removal order.
The European Commission also chairs the EU Internet Forum. It is a platform for EU Member States, European agencies like Europol and the Fundamental Rights Agency, tech companies, partners like the GIFCT and Tech Against Terrorism, and experts from civil society to come together and discuss approaches to combating terrorist and violent extremist content online, in a voluntary and cooperative format. Discussions take place under the Chatham House Rule, to enable trust and open information-sharing. Information about the EUIF’s objectives, rules of procedure, events and activities is published on its website.

**Good Practices**

- Establishes a framework for transparency and reporting that includes user notifications and reporting by tech companies and governments. Monitoring and implementation reports should also help enable the “big picture” to emerge.
- PERCI ensures systematic record keeping for relevant decisions and actions by all actors and should mean their reports are mutually consistent.
- Includes an evaluation of the impact of the regulation on security and human rights.
- EUIF provides a forum for multistakeholder discussion on emerging challenges and threats to develop joint action on a voluntary basis.

### 4: Transparency Reporting: EU Internet Referral Unit, Europol

The EU IRU has published annual transparency reports on a voluntary basis since 2017.

The reports describe EU IRU’s preventive work across three workstreams: monitoring and analysing terrorist propaganda; referring terrorist content and content related to migrant smuggling to online service providers for assessment against their terms of service; and work in public private partnerships e.g. in the EU Internet Forum and GIFCT Crisis Response Working Group.

The TCO Regulation specifies that the EU IRU (and Member States) retain the ability to refer content to online service providers for assessment and voluntary action under terms of service. The section of the EU IRU’s transparency report on referrals sets out the legal mandate for this aspect of the IRU’s work, provides a brief description of the processes and systems that the IRU uses – including any thematic and platform-specific referral action days held during the year - and gives the number of pieces of content assessed as being related to terrorism and violent extremism, and subsequently referred to an online service provider for consideration and action. There is no breakdown by e.g. platform or outcome.

See the [2021 report](#) for an example.

**Good Practices**

- Gives information on the mandate for and approach to ToS referrals.
- Describes where and how the EU IRU works with industry and other partners on issues of mutual concern e.g. crisis response, and research on emerging threats.

### 5: Transparency Reporting: Department of Internal Affairs, New Zealand

The Department of Internal Affairs set up its Digital Violent Extremism Team (DVET) in 2019, as part of the New Zealand Government’s response to the Christchurch terror attack. The Department has
published transparency reports for 2021 and 2022. The aim is to help the public understand how DVET works to prevent exposure to harmful and illegal content.

The reports contain information on the legal framework in New Zealand, as well as data on referrals against terms of service, take down notices for illegal content, and search warrants and productions orders issued for the purpose of investigating offences under the Films, Videos and Publications Classification Act. Data is presented in graphs with explanatory material. Case studies help readers understand how the team goes about its work, the kind of content it has dealt with during the year, and how domestic and international events have influenced the online environment. Additional information is provided on incident response, Official Information Act requests and media queries.

The Department has worked closely with Tech Against Terrorism’s and domestic partners (including the New Zealand Police and the Classifications Office) to develop these reports. Tech Against Terrorism reviews each report against the guidelines and makes public recommendations for further improvements.

**Good Practices**

- Transparency reports contain most of the categories and types of information and data recommended in Tech Against Terrorism’s guidelines.
- Data and information is presented in a way that helps different audiences interpret and understand it.
- The Department is also looking at innovations to make the data and information in the report even more readily accessible to key audiences, and for digital savvy youth.

### 6: Transparency Reporting, Ofcom, UK

Ofcom is the United Kingdom’s independent communications regulator, currently responsible for regulating UK-based video-sharing platforms and due to take on broader online safety powers following the passage of the Online Safety Bill. Ofcom works with platforms to ensure they have effective systems and processes in place to protect all users from relevant harmful material, as required by law. Ofcom does not make requests or referrals in relation to specific pieces of content. Rather, it focuses on platforms’ policies, systems and tools.

Ofcom published a transparency report on its first year of video-sharing regulation in October 2022. The report sets out the organisation’s mandate and approach, as well as a summary of its key findings about regulated services for the year and priorities for the year ahead. The report also includes more detail on Ofcom’s work with and assessment of the video-sharing platforms notified to it under the legislative regime, as well as with international partners. Ofcom has also published more detailed reports on specific incidents – such as the Buffalo, New York, terror attack in May 2022 – to analyse and address risks in the sector, as well as on its work with individual platforms to address particular harms.

With the passage of the Online Safety Bill in 2023, Ofcom will be required to set up a mandatory transparency framework for the regulatory regime, which will cover user-to-user platforms and search engines. Ofcom will consult on this framework but intends reporting requirements to be tailored to each specific regulated platform’s functionality, user base and unique characteristics.

**Good Practices**

- Ofcom publishes detailed information on its work with individual video-sharing platforms to improve online safety, as a basis for public trust and confidence.
• Ofcom is required to consult on the transparency framework under the Online Safety Bill. It intends that reporting requirements will be tailored to specific types of platforms and information and data will be made available in various formats, including static reports and real-time dashboards.
Supporting smaller platforms

Introduction
The Christchurch Call (the Call) was established in 2019 as a commitment by governments and Online Service Providers (OSPs) to eliminate Terrorist and Violent Extremist Content (TVEC) online. The Call rests on the conviction that a free, open, and secure internet offers extraordinary benefits to society and that respect for freedom of expression is fundamental; however, no one has the right to create and share TVEC online. The Call has since grown to over 120 governments, OSPs, partners, and civil society organisations, all working together towards the elimination of TVEC on the internet.

From the beginning, Call Leaders have recognised the important role that smaller OSPs play in preventing the proliferation of TVEC, particularly during crisis incidents. In 2019, they noted the awareness and capacity challenges that smaller OSPs face and, in response, supporters committed to empowering smaller OSPs as they build capacity to remove TVEC. Since 2019, the Christchurch Call has catalysed this support, with the Call Community building upon existing and new resources and tools that have revolutionised the support available to smaller OSPs that need assistance.

However, gaps remain that still challenge the ability of smaller OSPs to eliminate TVEC from their services. In the wake of global regulatory change, smaller OSPs are experiencing an increasing shortage of awareness and capacity to engage with the now-wide range of available resources and tools, and they often are unable to participate in effective crisis response mechanisms. Terrorists and violent extremists are exploiting these gaps to propagate their content as widely as possible.

Responding to these challenges, at their 2022 Summit Call Leaders committed to “ensuring shared industry tools remain at the leading edge of technology and are effective in responding to adversarial tactics and shifts”. This report seeks to analyse these challenges, map the current supports available to smaller OSPs, and consider what needs to be done to address identified gaps. It is based on desktop research and consultation with the Call Community and smaller OSPs between May and September 2023.

Definition of smaller OSPs
For the purposes this report, an OSP is the provider of any online service that facilitates the access to, input of, or upload of user-generated content. This definition generally excludes static websites without this functionality but can include those with comment sections (such as news sites), as well as file search engines, storage (including hosting), file sharing, social media, gaming, video, messaging, and some other services.

Referring to some OSPs as “smaller” can be misleading. Some have just a few dozen users whereas others can have millions. Some can have very few staff relative to users, while others can be better resourced. It is also not particularly useful from an analytical perspective. Consultations highlighted
Supporting Smaller Platforms

the diverse ways OSPs can be categorised, and how they categorise themselves, with reference to dimensions including Monthly Active Users (MAU), revenue, number of staff, number of dedicated trust and safety staff, nature of the service, and community culture. These different dimensions influence both the risk that a service will be exposed to TVEC and the OSP’s awareness, capacity, and willingness to address it.

For this report, a smaller OSP is any provider that lacks the awareness and/or capacity to eliminate TVEC on their service. *

Threat environment

It remains unclear how many smaller OSPs operate, out of the approximately 201 million active websites online. OFCOM estimates that approximately 100,000 services will be within the scope of the UK Online Safety Bill, although the number of these that are currently experiencing terrorist and violent extremist exploitation, and to what extent, is under-researched at this time.

The online ecosystem is also volatile, with services continually starting-up and shutting-down across the web, and decentralisation providing new portability of presence for users while creating new moderation risks. The geography and user base of these services is equally unclear, although is certainly in the hundreds of millions (potentially well over a billion), given the collective MAU of many services.

Amongst this dynamic and vibrant internet ecosystem, terrorists and violent extremists are increasingly exploiting smaller services to share TVEC. This is not a new issue: the United States Department of Homeland Security first identified in 2010 that jihadist organisations were utilising a multi-platform approach to sharing TVEC. Numerous academics and others have subsequently noted the tactics that terrorists and violent extremists use to propagate their content as widely as possible. This behaviour typically involves ‘outlinking’ (also called ‘dead-dropping’ or ‘signposting’) from beacon platforms (typically the largest OSPs, where content is quickly removed) to content stores (typically smaller OSPs) where the content resides and where rapid and effective content moderation is less likely to exist, aided by circumventors (such as VPNs) to avoid detection. The consequence of this behaviour is an increased exploitation of the awareness, capacity, and willingness of smaller OSPs, which also carries implications for the largest OSPs.

Smaller OPS are also exploited in crisis incidents, such as the 2022 Buffalo, NY, attack where the shooter exploited several smaller services to both plan and broadcast his attack. TVEC from this attack and the Christchurch massacre are continuously quickly removed by the largest OSPs but remains readily available on some smaller services where extremists exploit a lack of detection capabilities to re-post, discuss and share it and similar material. ‘Alt-tech’ platforms represent a subset of smaller OSPs dedicated to extreme or fringe content, usually with a lack of content moderation; these OSPs provide opportunities for terrorist and violent extremist exploitation, with little intervention.

* Willingness is also a key consideration, and unmoderated and alt-tech services are a priority outlined in the 2022 Christchurch Call Leaders’ Summit, however unwillingness raises a different set of issues and possible interventions, which the Call Community is to explore in depth in the Free, Open and Secure Internet Principle Working Group.
Landscape mapping

Regulatory change

Recent and upcoming government laws will require increased content moderation, risk assessments, transparency, and more from OSPs. For example, the European Union Terrorist Content Online Regulation requires compliance with removal orders within one hour, transparency reporting, and other measures such as preserving removed content and having an appeals and complaints mechanism. Australia’s Online Safety Act includes a risk assessment requirement. Singapore’s Online Safety Act includes the need to comply with blocking directions, and the United Kingdom’s Online Safety Bill will require OSPs to proactively remove illegal content; all of which include in scope smaller OSPs in some capacity. Evidently, regulators recognise that TVEC and other forms of illegal/harmful content can reside and be shared on smaller services, so regulatory responses must include them to be effective. This new regulatory burden may hinder innovation and competition because of the strenuous resources required to implement compliance measures. Most laws are designed to minimise effects on smaller OSPs, and some have also put in place specific supports (see below), but in a global ecosystem with a growing number of regulatory frameworks with their own compliance burdens, it is increasingly difficult for many smaller OSPs to possess the awareness and capacity necessary to adhere to all of them.

Existing resources and tools

There are numerous resources and tools available in the Call Community to support smaller OSPs counter terrorist and violent extremist exploitation of their services. Briefly, these include:

- **Tech Against Terrorism (TAT)**, whose efforts to support smaller OSPs include a membership program and Knowledge Sharing Platform that provide access to practice guides on defining terrorist content, transparency reporting benchmarks, global legislation (through the Online Regulation Series) and terrorist symbols and terms.

- The **Terrorist Content Analytics Platform (TCAP)**, developed by TAT with the support of the Canadian Government, which alerts participating smaller OSPs of URLs on their services that contain TVEC. Soon, the TCAP will also hash URLs for the GIFCT’s Hash Sharing Database and archive content for audit and research purposes.

- **TAT Europe**, the arm of TAT funded by the European Commission to uplift the capacities of smaller OSPs to comply with the Terrorist Content Online Regulation.

- The **Global Internet Forum to Counter Terrorism (GIFCT)**, whose membership provides access to cross-service information sharing, crisis response including the Hash Sharing Database, working groups, and workshops.

- **Hasher-Matcher-Actioner**, an open-source tool developed by Meta to assist smaller OSPs develop in-house capacity to label, hash and match harmful content.

- An upcoming tool from **Google Jigsaw** and TAT that will provide content moderation workflow capabilities to smaller OSPs that currently rely on manual moderation.

- The maturing commercial **safety tech industry**, comprising over 350 vendors that provide automated, behavioural-based detection of online harms, intelligence and threat awareness, age-assurance, regulatory compliance support, and more.

These tools, particularly those from TAT and the safety tech industry, have lowered the costs for smaller OSPs looking to eliminate TVEC from their services and have begun to fill the gap in
Supporting Smaller Platforms

resources and tools that existed in 2019. Nonetheless, a variety of challenges still face smaller OSPs, as will be presented below.

Gap analysis
Ideally, OSPs regardless of size should have the awareness, capacity, and willingness to eliminate TVEC from their services, while respecting human rights and playing their part in a free, open, and secure internet. Despite advancements in the tools and resources, there are remaining gaps in awareness and capacity that impact the ability of smaller OSPs to fully participate in online counter-terrorism efforts, which this section will explore.

As noted earlier, there are likely hundreds of thousands of smaller OSPs operating around the world, many of which could potentially be at risk from terrorist and violent extremist exploitation. It is important to note at the outset that many smaller OSPs likely have the awareness, capacity, and willingness to eliminate TVEC, even if they are not formally involved in multistakeholder or industry-led efforts. The Christchurch Call recognises that resources and tools should be available and accessible to all smaller OSPs that need or could benefit from them, regardless of their membership of specific initiatives.

Current gaps:

**Awareness:** The extent to which the majority of smaller OSPs are aware of the existing resources and tools available to them is unclear and there is limited research to rigorously evaluate this point. Anecdotally, however, we know from our consultations that levels of awareness are highly variable. Some are aware; others are not at all.

**Capacity:** Obtaining awareness of available resources and tools and then implementing them as an effective response to TVEC requires capacity, which is often lacking within smaller OSPs. Recently, cuts to trust and safety teams within smaller OSPs have further restricted their capacity to address specific risks on their services. A common theme heard in our consultations with smaller OSPs was their need to prioritise other kinds of illegal content, based on greater prevalence and/or harmfulness.

Given the significant changes – regulatory, technological, and commercial – that have impacted the OSP and TVEC landscape in the time since the launch of the Christchurch Call in 2019, the awareness and capacity of smaller OSPs remains in flux. Regulatory frameworks are still evolving, but combined with education and support programmes, they will meaningfully impact both the awareness and the cost-benefit analysis of acting against TVEC where it is a risk (i.e., capacity and willingness). At the same time, the efforts of Call supporters like TAT, Meta, Google and others, and the flourishing safety tech sector, means there are new resources and tools available to address the diverse needs of smaller OSPs, and these are increasingly affordable to acquire and implement. It will take some time for the full effects of these changes to materialise, and the consequent impact of TVEC on smaller OSPs to be fully understood.

In the meantime, the Call Community should continue to support smaller OSPs to improve awareness and build capacity to address TVEC – issues we know they face – with a focus on:

- **Increasing awareness amongst smaller OSPs of existing resources and tools to eliminate TVEC.** While there are now numerous resources and tools available to smaller OSPs, a complex regulatory landscape and competing online harm priorities mean that only some
smaller OSPs have a full understanding of what TVEC is, why it is important to eliminate it, why they are required to, and what exists to help them to do. The Call Community should undertake communication efforts, perhaps bundled with the campaigns of regulators who are aiming to increase compliance with their laws, by presenting existing resources and tools as a tangible way OSPs can demonstrate their online safety efforts.

- **Identifying, encouraging, and supporting smaller OSPs to avail themselves of these resources and tools.** While initiatives such as the GIFCT deliver valuable support and industry information sharing capabilities, strenuous membership criteria mean that these sorts of resources are an option for only the most motivated of smaller OSPs. Lower-barrier options, such as those from regulators, TAT, and the commercial safety tech market, are best-suited for smaller OSPs who are lacking in capacity. The Call Community should use its knowledge of which smaller OSPs are at the greatest risk of exploitation and use all available levers to improve their uptake of existing resources and tools.

- **Integrating smaller OSPs into effective crisis response.** One smaller OSP we consulted said that a terrorist crisis incident on their service is what “keeps me awake at night”, but crisis integration for smaller OSPs is currently limited to the GIFCT, whose membership criteria is beyond the capacity of most. TAT is currently developing a new tiered alerting framework and incident response policy for the TCAP, linked to existing crisis response protocols. Should TAT obtain funding to make the TCAP’s crisis functions available 24/7, it would create a globally available, low-barrier entry for smaller OSPs wishing to quickly become part of the crisis response network. The Call Community should consider support for the TCAP and other effective crisis response solutions as a priority to aide with the integration of smaller OSPs into effective crisis response.

- **Ensuring that solutions are transparent and respect human rights.** As obligations to eliminate TVEC increase for smaller OSPs, an increased take-up of resources and tools is encouraged (as above), but with this comes the potential for risks for human rights. We heard some concerns about privacy and freedom of speech implications of commercial safety tech, although some vendors offer assurances and mitigations for both. There are resources available, such as the TAT Transparency Guidelines and Human Rights Toolkit, that aim to uplift OSP awareness of human rights. The Call Community should consider how it can raise awareness and incentivise demand and supply for transparent, human-rights-respecting solutions for smaller OSPs.

- **Considering the impact of new and emerging technologies.** New tech such as generative artificial intelligence, decentralised services, blockchain and immersive technologies are sometimes presented as the next generation of the internet. Start-up and smaller OSPs lead the charge in the adoption of new tech, and often face the burden of designing and implementing content moderation in entirely new ways. For example, research points to the challenges decentralisation provides to eliminating TVEC. Generative AI poses its own challenges, but also presents opportunities for the automated detection of TVEC, particularly by the commercial safety tech sector. The Call Community is already actively considering the impact of new tech and should actively consider the unique challenges faced by smaller OSPs.

The role of willingness, as it relates to smaller OSPs, is also an important gap given the rise of alt-tech and far-right services. This topic will be explored in depth in the Call’s Free, Open and Secure Internet Principle Working Group.
Conclusion

The Call Community has catalysed the resources and tools available to smaller OSPs since 2019, with Call supporters and partners investing to ensure that services of any size can eliminate TVEC. However, challenges remain, and the Call Community should focus on supporting the awareness and capacity of smaller OSPs to engage with and effectively use these resources and tools, including supporting smaller OSPs to participate in crisis response. The Call can fill these gaps by building on the efforts already underway in the Community, with consideration paid to human rights and the impact of new tech, to support a free, open, and secure internet.

Thanks

Parts of this report are based on a doctoral dissertation34 by Amy-Louise Watkin, in which she proposes a framework to regulate terrorist content online. In her framework, Dr. Watkin suggests that the awareness, capacity, and willingness of OSPs are the key obstacles to successful intervention against TVEC. We thank Dr. Watkin for her research and the time she spent providing advice to us on this topic.

We also thank all Call supporters and partners and the OSPs that participated in the consultation for this report. A particular thanks goes to the smaller OSPs who invested time speaking to us despite their limited time and staff.

Endnotes

1 Christchurch Call. (2019). The Christchurch Call to Action: To eliminate terrorist and violent extremist content online.
3 Siteefy. (2023). How many websites are there in the world?


21 Ware, J. (2020). *Testament to murder: The violent far-right’s increasing use of terrorist manifestos*. International Centre for Counter-Terrorism.


25 Tech Against Terrorism Europe. (2023). *European regulation on Terrorist Content Online (TCO)*.


29 Tech Against Terrorism. (2023, May 5). *Terrorist Content Analytics Platform launches in Canada* [Press release].

30 PUBLIC. (2023). *The international state of safety tech*.


Misogynistic Pathways to Radicalisation:
Recommended Measures for Platforms to Assess and Mitigate Online Gender-Based Violence

Sara Bundtzen
About the Digital Policy Lab

The Digital Policy Lab (DPL) is an inter-governmental working group focused on charting the policy path forward to prevent and counter the spread of disinformation, hate speech, extremist and terrorist content online. It comprises representatives of relevant ministries and regulatory bodies from liberal democracies. The DPL aims to foster inter-governmental exchange, provide policymakers and regulators with access to sector-leading expertise and research, and build an international community of practice around key challenges in the digital policy space. We thank the German Federal Foreign Office for their support for this project.

About this Paper

As part of the DPL, the Institute for Strategic Dialogue (ISD) organised working group meetings on the topic of online gender-based violence between May and June 2023. The working group comprised DPL members representing national ministries and regulators from Australia, Canada, Germany, New Zealand, the UK, and the US. Members of the Global Partnership for Action on Gender-Based Online Harassment and Abuse (Global Partnership) and the Christchurch Call’s multistakeholder Community joined the working group and contributed to this paper. Participants also included representatives from civil society, academia and industry.

While participants contributed to this publication, the views expressed in this paper do not necessarily reflect the views of all participants or any governments involved in this project.

About the Author

Sara Bundtzen is an Analyst at ISD, where she studies the spread of information manipulation by state and non-state political actors in multilingual online environments. As part of the Digital Policy Lab (DPL), Sara informs ISD’s advisory work and analyses proposed pathways toward countering disinformation, influence campaigns, hate speech, and extremist content. Sara previously worked at the Federal Ministry of Defence and NATO HQ.

Contributing editor

Helena Schwertheim is a Senior Digital Policy and Research Manager at ISD Germany, where she leads the Digital Policy Lab (DPL). Helena is interested in government and multistakeholder responses to the digital threats posed to democracy by disinformation, hate speech and extremism. Previously, Helena managed digital policy and research projects at Democracy Reporting International, and she has experience in risk and political analysis in international organizations and think tanks, including at the UN World Food Programme, and International IDEA.

Editorial responsibility

Huberta von Voss, Executive Director, ISD Germany & Henry Tuck, Head of Digital Policy, ISD

Acknowledgements

We would like to thank all participants of the working group, including experts from governments, civil society, academia, and industry, for their contributions. We would like to give special thanks to the speakers as well as contributors to this paper for providing valuable insights and feedback: Elisabeth Brown (Christchurch Call Unit, Department of the Prime Minister and Cabinet (DPMC), New Zealand), Cailin Crockett (National Security Council / Gender Policy Council, The White House, US), Stella Ivory (Christchurch Call Unit, DPMC, New Zealand), Clara Martiny (ISD), David Reid (Christchurch Call Unit, DPMC, New Zealand), Sara Soleymani (Ministry for Women, New Zealand), Caroline Sinders (Convocation Research + Design), and Professor Lorna Woods (University of Essex, UK). We would like to thank ISD colleagues Milo Comerford, Cooper Gatewood, Melanie Smith, Dr. Tim Squirrell, and Henry Tuck who reviewed this paper.

Copyright © Institute for Strategic Dialogue (2023). The Institute for Strategic Dialogue (gGmbH) is registered with the Local Court of Berlin-Charlottenburg (HRB 207 328B). The Executive Director is Huberta von Voss. The address is: PO Box 80647, 10006 Berlin. All rights reserved.

www.isdgermany.org
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Executive Summary</td>
<td>4</td>
</tr>
<tr>
<td>Glossary</td>
<td>8</td>
</tr>
<tr>
<td>Introduction</td>
<td>9</td>
</tr>
<tr>
<td>Trends in online gender-based violence</td>
<td>10</td>
</tr>
<tr>
<td>Actors: OGBV as a vector for radicalisation and violent extremism</td>
<td>10</td>
</tr>
<tr>
<td>Tactics: OGBV as a continuum of violence</td>
<td>12</td>
</tr>
<tr>
<td>Harms: impact of OGBV on individuals and society</td>
<td>14</td>
</tr>
<tr>
<td>Assessment of platform policies and enforcement</td>
<td>15</td>
</tr>
<tr>
<td>Platform design and systems: Risk of reproducing and amplifying OGBV</td>
<td>16</td>
</tr>
<tr>
<td>Response measures: Risk assessment and mitigation of OGBV</td>
<td>18</td>
</tr>
<tr>
<td>Enable API access to platform data and develop standardised transparency reporting</td>
<td>18</td>
</tr>
<tr>
<td>Apply a victim-survivor-centred Safety and Privacy by Design approach</td>
<td>19</td>
</tr>
<tr>
<td>Enhance cross-platform cooperation and information sharing</td>
<td>20</td>
</tr>
<tr>
<td>Review and update content moderation policies, processes, and systems</td>
<td>20</td>
</tr>
<tr>
<td>Audit and mitigate misogyny in AI-based systems</td>
<td>21</td>
</tr>
<tr>
<td>Conclusion</td>
<td>23</td>
</tr>
<tr>
<td>Endnotes</td>
<td>24</td>
</tr>
</tbody>
</table>
Executive Summary

This paper reviews online gender-based violence (OGBV) as existing within a continuum of (online and offline) violence, emphasising the connections with different extremist ideologies, including the dissemination of terrorist and violent extremist content (TVEC). It aims to prioritise a gender perspective in responding to TVEC so that social media platforms can better intervene in and mitigate misogynistic pathways to radicalisation that can begin (or be reinforced) online.

Given the scope of this review, focusing on platforms rather than broader forms of digital technologies, the paper uses the terminology OGBV in place of technology-facilitated gender-based violence (TFGBV).

The discussion asserts that the root causes of gender-based hate, misogyny, and other intersecting forms of identity-based hate and violence mirror a broader societal challenge that cannot be addressed or fixed by platforms alone. It thereby recognises that the mitigation of OGBV and online pathways to radicalisation requires a whole-of-society and whole-of-government approach. Whilst there are steps that governments and civil society can and should take, such as overseeing and enforcing emerging regulatory frameworks and voluntary commitments, this paper and its recommendations emphasise the role and actions of platforms.

Outlining the impact of OGBV at micro (individual) and macro (societal) levels, this paper considers how OGBV can be a vector for radicalisation, and is motivated by misogyny, which also pervades terrorist and violent extremist ideologies. The paper concentrates on the role platforms can play in exacerbating the risks of OGBV, evaluating platform policies, content moderation practices, user interface design and algorithmic recommender systems.

The discussion considers OGBV as any form of violence, including dehumanising language, directed against persons based on their gender identities or expressions, with intersecting protected characteristics such as (but not limited to) race, indigeneity, religion, sexual identity, class, or disability increasing the risks of experiencing OGBV. The paper recognises that women and LGBTQ+ people experience OGBV disproportionately. It situates OGBV as inherently linked to longstanding patriarchal gender norms, with misogyny functioning as an ideological link across a continuum of violence and as a vector across different extremist ideologies.

In this context, the paper asserts that researching and mitigating the risks of OGBV can enable earlier warning of and intervention in misogynistic pathways to different forms of violent extremism. Reiterating that any mitigation of risks must come in support of users’ fundamental rights, including their right to privacy and freedom of expression, the paper proposes the following key recommendations.

Key recommendations

**Enable API access to publicly available data for public interest research:**

- The systematic collection of publicly available data via access to Application Programming Interfaces (APIs) can help complement digital ethnographic (and other) research methods, by filling in data gaps for the purpose of public interest research.
- Platforms should enable access to continuous, real-time or near real-time, and searchable APIs to allow vetted researchers to study the evolving tactics and forms of OGBV, as well as the links between online misogyny, radicalisation pathways and violent extremism. For example, such access could support longitudinal studies of in-group gender norms and behaviours over time, across different extremist ideologies and across platforms.
- Vetting processes of researchers should be inclusive to enable interdisciplinary research, involving a range of disciplines such as Computational Linguistics, Critical Terrorism Studies, and Critical Studies on Men and Masculinities, as well as recognising the value of comparative research across different ideologies, local contexts, and languages.
- While API access requires some form of vetting to prevent malicious or commercial uses, access should be free or at a nominal cost for researchers. Higher costs risk a de-facto inability to access data, or inequity among less well-resourced researchers.
Develop gender-disaggregated and standardised transparency reporting:

- Transparency reporting by platforms should enable external researchers to track and scrutinise the scope and scale of OGBV and the enforcement of community guidelines over time.

- Platforms should develop enforcement reports to include gender-disaggregated data, referring to statistical data in relation to community guideline violations. For example, platforms’ processing of, and reporting on hate speech violations should include data on whether the violation was on the basis of gender and other intersecting protected characteristics, to allow intersectional analysis of the motivations driving OGBV.

- Platforms should work towards standardisation of transparency reporting through the development of a set of common metrics and content categories, whenever possible, to allow for comparison and tracking of policy violations across platforms. For example, standardised reporting should disclose the proportion of image-based versus text-based content that violated hate speech policies.

- A cross-platform effort to standardise reporting could coordinate and align with work by UN Women to develop a statistical framework for TFGBV and the UN Special Rapporteur on Freedom of Expression to develop a common definition for gendered disinformation.

- Platforms should consult and collaborate with GBV and feminist advocates, scholars, and victims-survivors with lived experience when developing the methodology of transparency reports or any internal research (e.g., when conducting user surveys). At a minimum, platforms should be transparent about the methodology of their reporting (and any changes thereof).

Apply a victim-survivor-centred Safety and Privacy by Design approach:

- Taking a victim-survivor-centred perspective, the development of user interfaces and tools should apply a gender and trauma-informed lens throughout all stages.

- Platforms should adopt proactive measures that support user agency with tools that protect their privacy and reduce exposure to OGBV; reactive measures that allow efficient user reporting (where possible, across platforms); and accountability measures that deter and sanction perpetrators appropriately.

- Data privacy and security should be embedded not only via accessible and transparent settings, but also in platforms’ policies to moderate and mitigate the use of personal data for OGBV (e.g., to prevent doxing or sharing of intimate images without consent).

- Content moderation tools developed by industry such as Google’s Perspective API should be continuously tested and scrutinised, taking into account a victim-survivor-centred perspective. Such efforts should be part of a cross-sector and multistakeholder dialogue.

Enhance cross-platform cooperation and information sharing of OGBV incidents (including actors and tactics):

- Platforms should develop and operate exchange channels between relevant teams, including content moderation teams, to proactively share information about OGBV incidents, including cross-platform harassment (such as relevant information about perpetrators using accounts across platforms). This is important for understanding the scope and scale of OGBV, but also to coordinate cross-platform responses and mitigation actions, where appropriate.

- Platforms should share information about user reports, where appropriate, and develop interoperable reporting mechanisms, where possible, to support user agency.

- Existing cross-platform coordination such as the Christchurch Call Crisis Response Protocol and the Global Internet Forum to Counter Terrorism should review how OGBV is relevant to their mandates and adapt their scope and mechanisms appropriately.

- Cross-platform knowledge exchange should further build on and improve existing content moderation tools, including through regular assessments and reporting about the efficiency and impact of these tools.
Review content moderation policies, processes, and systems to acknowledge the continuum of violence and misogyny as a vector for violent extremism:

- Content moderation should account for the continuum of violence and recognise misogyny as a gateway and early warning sign of different extremist ideologies.

- Platforms should review and update hate speech and TVEC policies to recognise how misogynistic beliefs pervade ideological pathways to extremism. This includes considerations of how misogyny can be an ideology that encourages violent extremism and justifies violence towards women and LGBTQ+ people.

- Review processes should include the perspectives of women, the LGBTQ+ community, and victims-survivors. Such processes should also be inclusive of other protected characteristics, including race and religion, noting that GBV towards racialised communities often comes through the vehicle of racist and dehumanising language.

- Content moderation should account for veiled and coded misogynistic content, including contextual image-based content, as well as the multilingual, cross-cultural contexts of online spaces. For example, platforms could develop lexicons of words and phrases in cooperation with local organisations. Such efforts should be trauma-informed.

- The use of Artificial Intelligence (AI)-based systems for the purpose of detecting and moderating misogynistic content needs to be complemented by human oversight to allow for nuanced approaches that recognise the role of subtle and veiled misogyny, while also preventing false positives.

Apply intersectional feminist knowledge in risk assessments of AI-based systems:

- Platforms should incorporate gender analysis and feminist methodology when assessing the risks of algorithms and machine-learning (ML) models embedded in their services. This approach is useful for understanding how structural gender inequalities and patriarchal gender norms can be reproduced and amplified by AI-based systems.

For example, platforms should review and update their recommendation guidelines (e.g., guidelines for content lowered in feeds) in alignment with a review of their community guidelines.

- Platforms should adopt victim-survivor-centric design processes from the ideation, conceptualisation, developing, testing, and scaling of new features or any changes to existing ones.

- Platforms should ensure that relevant teams (such as those designing, testing, and evaluating algorithms) are diverse and trained on how to conduct gender analysis to detect and mitigate biases and discriminatory patterns in their systems.

Strengthen and encourage multistakeholder dialogue and collaboration:

- As part of a broader good-faith effort, platforms should contribute to a trusted environment that supports exchange between stakeholders, including policymakers, government agencies, civil society, academia, product developers and software engineers.

- Multistakeholder exchange should actively seek intersectional perspectives, including those of victims-survivors of OGBV.

- Regular exchange could focus on testing and enhancing methods for assessing and mitigating OGBV. For example, stakeholders could discuss interoperable user reporting, content moderation tools, and algorithmic pathways.

- Multistakeholder collaborations should encourage the exchange of resources. Platforms and partners could consult the following resources for additional evidence and recommendations:

  - “Technology-facilitated gender-based violence: preliminary landscape analysis” by the Global Partnership for Action on Gender-Based Online Harassment and Abuse;
Recommended Measures for Platforms to Assess and Mitigate Online Gender-Based Violence

- “Technology Companies Must Make Platforms Safer for Women in Politics”, “Interventions to End Online Violence Against Women in Politics” and “Landscape Tracker” by the National Democratic Institute (NDI);

- “The Chilling: A global study of online violence against women journalists” by the International Center for Journalists (ICF)/UNESCO;

- “Guidance on the Safe and Ethical Use of Technology to Address Gender-based Violence and Harmful Practices” by the UN Population Fund (UNFPA);

- “Measuring technology-facilitated gender-based violence. A discussion paper” by the UN Population Fund (UNFPA);

- “Report on freedom of expression and the gender dimensions of disinformation” by the UN Special Rapporteur on the promotion and protection of freedom of opinion and expression;

- “Technology-facilitated violence against women: Taking stock of evidence and data collection” by UN Women;

- “Violence Against Women and Girls (VAWG) Code of Practice” by The End Violence Against Women Coalition, Glitch, Refuge, Carnegie UK, NSPCC, 5Rights, Professor Clare McGlynn and Professor Lorna Woods.
Glossary

Application Programming Interfaces (APIs) are software intermediaries that allow two applications to communicate with each other. APIs have a huge range of uses, but in the context of this Explainer, they allow researchers to access certain types of data from some online platforms via requests. As an intermediary, APIs also provide an additional layer of security by not allowing direct access to data, alongside logging, managing and controlling the volume and frequency of requests.

Extremism is the advocacy of a system of belief that claims the superiority and dominance of one identity-based ‘in-group’ over all ‘out-groups’. It propagates a dehumanising, ‘othering’ mind-set incompatible with pluralism and the universal application of human rights. According to ISD’s definition, extremism can manifest through violence and the targeting of hate towards groups on the basis of their identity, as well as more gradualist supremacist social or political projects that undermine human rights, democratic institutions and civic culture. It is important to place OGBV on the spectrum of extremism as misogynistic content is used in radicalisation processes, and can incite and translate to offline violence. Violent extremism is understood in this context as a specific violent manifestation of the wider phenomenon of extremism.

Gender is an individual’s internal sense of being a woman, a man, neither of these, both or somewhere along a spectrum. It describes socially constructed roles for women and men, and is an acquired identity that is learned, changes over time and varies widely within and between cultures. Gender norms or gender stereotypes are “generalised views or preconceptions about attributes or characteristics, or the roles that are or ought to be possessed by, or performed by, women and men.” They are often framed in a binary that overlooks the lived experience and richness of gender-diverse people, while also being trans exclusionary. In contrast, sex is assigned at birth based on the physical appearance associated with being female or male.

Gender-based violence (GBV) refers to “violence directed against a person because of that person’s gender or violence that affects persons of a particular gender disproportionately.” Women and LGBTQ+ community, including transgender and gender-diverse people, experience disproportionate rates of GBV.

Male supremacy is a “hateful ideology rooted in the belief of the innate superiority of cisgender men and their right to subjugate women [and trans and gender-diverse people].” It is linked to hegemonic masculinity, which structures patriarchy and describes the “legitimation of unequal gender relations.”

The Manosphere is an umbrella term that refers to several interconnected misogynistic communities online. It encompasses multiple types and severities of misogyny with varying expressions of violence – from broader male supremacist discourse to Pick Up Artists, Men’s Rights Activists (MRAs), Men Going Their Own Way (MGTOW), and involuntary celibates (incels).

Misogyny operates to uphold a patriarchal social order, policing gender norms to ensure that women and marginalised gender identities conform. It works to justify violence if these norms are deviated from. Misogyny includes what might be considered a type of deeply held prejudice towards women and marginalised gender identities and intersects closely with racism, antisemitism, Islamophobia, ableism, and anti-LGBTQ+ hate. Misogyny thereby operates alongside other intersecting forms of discrimination, including misogyny targeted at transwomen (transmisogyny) and the specific form of hatred Black women face (misogynoir). It is often hidden within different forms of violent extremist ideologies. It is also a motivating ideology in itself, separate from other types of extremist ideologies.

Online gender-based violence (OGBV) can be described as a subset of technology-facilitated gender-based violence (TFGBV), which refers to any “act that is committed, assisted, aggravated, or amplified by the use of information communication technologies or other digital tools, that results in or is likely to result in physical, sexual, psychological, social, political, or economic harm, or other infringements of rights and freedoms.” The phenomenon is also referred to as technology-facilitated violence against women (TFVAW), noting that VAW can be substituted with GBV, whilst maintaining the common definition describing the phenomenon.

Radicalisation is a term used in this context to describe the process by which an individual adopts an extremist ideology (defined above), which may (or may not) enable acts of violent extremism or terrorism. In the literature on terrorism and violent extremism specifically, a frequent distinction is made between cognitive radicalisation (adopting extremist beliefs) and behavioural radicalisation (the process leading up to violent behaviour).
Introduction

Gender-based violence (GBV) in public and private life is a global challenge that has been increasingly connected to and amplified by the online spaces of social media platforms, messaging services and other communications technologies. It reflects the manifestation and amplification of unequal power relationships that stem from patriarchal gender norms, which can be directed at all genders, but most often towards women and LGBTQ+ people. It also intersects with other forms of identity-based violence such as (but not limited to) racism, Islamophobia, and antisemitism.

While social media platforms can help empower feminist movements, for example, by bringing greater visibility and attention to women’s and LGBTQ+ communities’ rights issues, the current online environment can enable and reinforce misogynistic and anti-LGBTQ+ content. Further, OGBV disproportionately affects women in public life, including activists and human rights defenders, politicians, and journalists causing a ‘chilling effect’ on equal civic and political participation – with gendered and sexualised mis- and disinformation also being used as deliberate tactics by (both non-state and state) anti-democratic actors.

The level of response by platforms to address misogyny on their services varies, and some have taken commendable actions. However, so far, no platform has identified and taken sufficient steps to effectively address the individual and societal risks emanating from OGBV.

While online manifestations of GBV have distinct features, they belong to a “continuum of multiple, recurring and interrelated forms of GBV.” OGBV enforces and amplifies the patriarchal order with tools from across a tactical spectrum, ranging from legal but harmful behaviour to terrorism and violent extremism. GBV manifests online, while the reproduction and amplification of misogyny online can lead to offline violence – ranging from intimate partner violence, physical attacks against female journalists to mass violence.

In this context, it is important to recognise that OGBV occurs within an ecosystem characterised by a gender digital divide that is rooted in structural gender inequalities, in which those who design – and, in some countries or regions, access and use – communication technologies are disproportionately male. Recognising the intricate dynamics of online and offline GBV, this paper elucidates the connection of OGBV with extremist ideologies and violent extremism, with a focus on evaluating and addressing the role and actions of platforms.

A range of multilateral fora including UN Women, the World Health Organisation (WHO), and the UN Population Fund (UNFPA), as well as multistakeholder initiatives such as the Global Partnership recognise the need for action to address the role of platforms in enabling and exacerbating OGBV. The Christchurch Call has recognised the need to deepen and explain the evidence base on the links between misogyny and TVEC.

Drawing on discussions with government, industry and civil society stakeholders, this paper reviews the trends and multi-level impacts of OGBV, emphasising the multifaceted relationship with violent extremism. Based on this review, the paper proposes risk assessment and mitigation measures for platforms to respond to misogynistic content and behaviour on their services.
OGBV has tangible and measurable offline impacts, and offline harms can be extended and amplified online. In many cases, the victim-survivor knows the perpetrator, who is often a current or former partner, relative, co-worker, or friend. Recognising the commonality of gendered power relations as elements of both intimate partner violence and extremist ideologies, the following section examines how misogyny, including gendered and sexualised motives and attitudes, overlaps with or becomes a vector for violent extremism. It acknowledges the need to consider the intersectionality of OGBV with other forms of violence such as racism and Islamophobia, as well as the relationship of online/offline manifestations of violent extremism.

**Actors: OGBV as a vector for radicalisation and violent extremism**

In recent years, there has been an increasing focus on gender dynamics in the context of research studying online radicalisation and the dissemination of TVEC across different extremist ideologies. Some scholars have noted the need for further recognition of misogyny as an ideological vector for radicalisation in Preventing and Countering Violent Extremism (P/CVE) programming and policy.

A growing body of research focuses on the relationship between far-right extremism, misogynistic ideology and the Manosphere; the latter being a loose network of misogynistic online communities that seek to enforce male supremacy and patriarchal gender norms. For example, the Southern Poverty Law Center (SPLC) explains that male supremacists “are fixated on rigid gender roles and vilify any deviation from their strict gender dichotomy,” describing male supremacy as a “powerful undercurrent for white supremacy, and its tenets undergird much of the contemporary far right.” Misogynistic violence has manifested in physical attacks on women, with misogynistic motivations also intersecting with racist and xenophobic sentiments. For example, the perpetrator of the spa shootings in Atlanta, Georgia in 2021, killing six women of Asian descent, displayed “gross misrepresentations of hypersexualized Asian women.”

However, the relationship between misogynistic groups such as ‘involuntarily celibates’ (incels) – a subset of the Manosphere who blame women and society for their lack of romantic success — and other supremacist ideologies is complex and multifaceted. For example, incels possess a unique perspective on race and ethnicity that differs from far-right groups, involving a perceived racial hierarchy in the dating sphere favouring white men, which they attribute to female choices in selecting sexual partners (rather than actively endorsing it). Far-right groups meanwhile drive a more racially supremacist vision, looking to enforce racialised sexual boundaries to maintain in-group homogeneity. In recognition of these nuances, there has been cross-pollination as both, incels and far-right groups, share a misogynistic ideology and antifeminist sentiments. ISD research notes that some incels explicitly identify with racially or ethnically motivated violent extremism (REMVE), for example, by labelling themselves ‘stormcels’ in reference to Stormfront, a notorious white supremacist website.

In this context, P/CVE policy and programming has lacked a focus on the potentially violent extremist outcomes of misogynistic ideology. Notably, recent scholarly debates have concentrated on whether examples of incel associated violence should be understood as constituting terrorism. While incels are certainly political in nature with a core ethos geared towards subjugating and repressing a group of people, there is no consensus on whether violence by this group should be considered primarily ideological, or alternatively nihilistic. In this context new legal precedents are being set, with the Ontario Superior Court of Justice recently determining an incel-motivated murder amounted to terrorist activity. In such instances, notions of ‘lone-wolf’ actors can be a misnomer. While perpetrators of GBV, both online and offline, might not affiliate to a particular group, this may be due to the nature of misogynistic ideology as diffuse, networked and pervasive, a phenomenon related to the wider challenge of ‘post-organisational’ extremism.

Misogyny — like antisemitism — often serves as a unifying core feature of different extremist ideologies. In the context of promoting hetero-normative gender norms and identities, the connections between violent extremism and misogyny showcase parallels between militant masculinity in different ideologies, notably in far-right and Islamist extremist groups. Researchers note that both “equate manliness with the readiness...
to defend and it is not uncommon for overt or diffuse misogyny to serve as a motivating force for turning to the respective ideology.”

Examining violent Islamist extremist actors, researchers observed that ISIS had practiced “a militarised, masculinised, religious and genocidal nationalism within their ‘Islamic State’ when subjecting Yazidi women and girls and other minorities to GBV.” Both Islamist and far-right extremism “impose patriarchal gendered roles, binaries, hierarchies, and norms,” reiterating that male supremacist and misogynistic belief systems are present across a diverse ideological spectrum. Common to all is their misuse and exploitation of mainstream and ‘alternative’ platforms and messaging services, cutting across geographical locations and languages.

In addition, researchers highlight how antisemitism intersects with misogynistic beliefs. Evelyn Torton Beck explains that the ‘Jewish Princess’ stereotype “remodels the traditional antisemitic tropes onto a female form: she is materialistic, money-grabbing, manipulative, shallow, crafty and ostentatious.” Blyth Crawford notes that the neofascist militant accelerationist movement sees Jewish people “as influencing sexual politics in ways that are regarded as being ‘anti-family’ and therefore constitute a threat to the white race.”

Part of the ‘white genocide’ conspiracy theory, there are connections between antisemitic tropes and anti-LGBTQ+ fearmongering, involving gender-based hate speech. The Anti-Defamation League (ADL) asserts that the “alleged targeted promotion of LGBTQ+ identities and relationships is seen as a key element of Jews’ attempts to reduce reproduction rates among straight, cisgender white people.” Anti-LGBTQ+ tropes labelling the community as “pedophiles” or “groomers” often converge with the antisemitic canard that Jews prey upon non-Jews, especially non-Jewish children. This observed confluence of antisemitism, misogyny and anti-LGBTQ+ hate is perpetuated by a wide range of extremists with different ideological backgrounds.

Structural gender inequality and gender norms can also lead to internalised misogyny. An extreme example is the emergence of Tradwives as an influential online community, showcasing the reinforcing elements of far-right ideology, Christian Nationalism, white supremacy, and patriarchal gender norms. Tradwives embrace a highly hetero-normative rendition of the ‘wife and mother’ role, in opposition to feminism, reproductive rights, LGBTQIA+ rights, and gender equality. Researchers highlight that Tradwives “use their presence on social media to offer a powerful female in-group association” and “successfully infiltrated mainstream social media with their anti-globalist, anti-modern approach to life.” While women’s role in violent extremist ideologies and communities varies, research finds that women-only forums have also “served as gendered sites of ideological contestation,” where women are asserting “agency in their everyday practices despite otherwise constraining gendered ideological constructs.”

While extremist communities often use veiled and coded language to conceal and convey in-group culture, online misogyny can become widespread and popular. ISD finds that it is most impactful and prolific when different ideological groups participate in the spread of misogyny. Anti-drag and anti-LGBTQ+ activities, for example, are not limited to fringe groups, but have become a unifying concern for the far-right as well as localised activists, including certain parents’ rights groups, anti-vaccine or anti-lockdown groups, and Christian nationalists. Additionally, OGBV and misogyny has been highly associated with violent conspiracy movements. For example, ISD research on the online activities of QAnon supporters has shown how targeted hate, including violent misogynistic, racist, and anti-LGBTQ+ rhetoric, has become a particular concern for prominent women, who often found themselves on the receiving end of coordinated harassment. Finally, ISD research shows that a small group of actors can have considerable influence over the propagation of misogynistic content, including for example prominent influencers like Andrew Tate.

Misogynistic content can thereby serve as an ideological link across different extremist groups, with increased exposure to online misogyny risking a normalisation among users, especially among male users who use online spaces to socialise, network, and connect with others.
Tactics: OGBV as a continuum of violence

As stated, OGBV occurs within a continuum of violence, which recognises the complex and interlinked experiences of different forms of violence.

This section outlines prevalent forms of OGBV, reviewing a 2020 survey on ‘Measuring the prevalence of online violence against women’ conducted by the Economist Intelligence Unit (EIU) as well as other relevant research, including by PEN America, the Global Partnership, and UNESCO/International Center for Journalists (ICFJ). It is not intended to be exhaustive, but to indicate the myriad forms of OGBV. Importantly, these forms and behaviours are frequently observed in combination and across multiple platforms.

• **Online gendered or sexualised mis- and disinformation** refers to “a subset of online gendered abuse that uses false or misleading gender and sex-based narratives against women [and trans and gender-diverse people], often with some degree of coordination, aimed at deterring women [and trans and gender-diverse people] from participating in the public sphere.” It may involve defamatory comments that intend to harm a person’s reputation. A combination of false information with the publication of factual, decontextualised and misrepresentative information is often the most harmful. Gendered and sexualised mis- and disinformation often uses coded and veiled language as well as iterative, context-based visual and textual memes.

• **Online harassment** encompasses a wide range of unwanted or negative contact that is used to create an intimidating, annoying, frightening, or even hostile environment. It can involve long-lasting coordinated narrative framing, sharing of target lists, and brigading across platforms. It may also be in the form of a single comment or one-off incident. It is often gendered or sexualised in nature.

• **Online gender-based hate speech** attacks or humiliates persons based on their gender identities and expressions, with intersecting identity factors such as (but not limited to) sexual identity, ethnicity, race, religion, or disability increasing risks of becoming a target of hate speech. It can range from dehumanising and derogatory language to threats and incitements of violence.

• **Online impersonation** refers to wrongfully obtaining and using another person’s personal data in some way that involves fraud or deception. Gendered examples include creating fake accounts to groom and recruit girls and women into sex trafficking.

• **Stalking and monitoring** involve the misuse of technology, such as installing commercial stalkerware on a device. Stalking and monitoring is often repeated, and can be an extension of intimate partner violence.

• **Astroturfing** refers to the deceptive practice of dissemination or amplification of content that appears to arise organically at the grassroots level, but is actually coordinated by an individual, interest group, political party, or organisation. Astroturfing may be part of networked harassment, which involves tactics such as trolling (purposely upsetting or disrupting online events, debates or hashtags) and coordinated flagging (falsely reporting users to get them de-platformed).

• **Image-based sexual abuse** involves the creation, distribution, sharing or threat of sharing intimate images or videos of a person without their consent. It includes a diversity of behaviours such as sexual extortion (when a person has, or claims to have, a sexual image of another person and uses it to coerce them into doing something they do not want to do), documentation or broadcasting of sexual violence posted on social media, texted among peers, sold or traded, resulting in an additional form of sexual violence against the victim-survivor, and the use of generative AI to construct deepfakes, including artificial images or videos that resemble actual photographs or videotapes.

• **Doxing** involves retrieving and publishing of personal or identifying information (e.g., addresses, phone numbers, emails, partners’ or children’s names) without permission — often with a malign intent to show up at the workplace or home, or to make negative or unwanted contact.
• **Threats of offline violence** such as rape and death threats, or incitement to physical violence. Women journalists, academics, politicians and human rights defenders often face violent threats, which are gendered and sexualised, particularly if they are speaking or writing about equality issues or male-dominated topics. A global study conducted by IFCJ mapped the vicious circular trajectory of online violence, highlighting that “digital attacks can fuel offline violence, while offline abuse by prominent figures can trigger online pile-ons.”

These forms of OGBV occur across platforms, often simultaneously and in a coordinated manner. Astroturfing and networked harassment tactics misuse platforms to facilitate wider reach of misogynistic content as well as the networking and in-group building of perpetrators who might otherwise be isolated from one another. For example, incel forums are spread across Reddit and 4chan as well as gaming forums like Discord or dedicated websites, which reinforce in-group, community and belonging. While this type of cross-platform misogynistic behaviour and networking creates additional risks for users, it also reiterates the challenge of understanding how vulnerable individuals become radicalised and how a healthy online environment can help prevent this from happening, including tailored interventions that address individual grievances.

Moreover, coordinated harassment campaigns often take advantage of online conversations surrounding trending topics, which may involve the use of abusive hashtags, to spread misogynistic content. This can also manifest in the coordinated harassment of an individual across multiple platforms. In turn, coordinated online harassment raises challenges for the tracking and reporting of OGBV, often putting the onus on victims-survivors, and illustrating the need for victim-survivor-centric coordination and collaboration between platforms.

Finally, the inherently global reach of many platforms expands the online misogynistic influences in radicalisation pathways, contributing to a perpetuation of online cultures of extremist beliefs. For example, studies have shown that occasional encounters with extremist content are experienced by 40% to 50% of younger individuals. This creates constant opportunities for the initiation of radicalisation processes within large populations.
Harms: impact of OGBV on individuals and society

This section outlines the impact of OGBV at both the micro (individual) and macro (societal) level, recognising the range of harms, including the risks to private as well as public safety.

- **Psychological harms:** Research shows that OGBV can leave victims-survivors with serious psychological harms, mental or emotional stress, as well as symptoms of post-traumatic stress disorder, particularly when the abuse is frequent. Cumulative effects of offline and online violence can also lead to self-harm, depression, and suicide.\(^7^6\)

- **Threats to reproductive health:** Misinformation about abortion and reproductive rights can cause gendered harm as it undermines access to correct information about health care and promotes unsafe alternatives or unproven medication.\(^7^7\)

- **Privacy invasions:** Once personal information is released online (for example via doxing), it can be difficult, if not impossible, to retrieve or remove. This also create risks of future invasions given that the information remains permanently on the Internet or stored on another person’s device. This negatively impacts the right to privacy of targeted persons.

- **Economic and material harm:** The term ‘economic vandalism’ highlights the economic costs caused by OGBV, for example, due to missed work opportunities, decreased productivity, and retreating from the Internet.\(^7^8\)

- **Exacerbating structural gender inequality:** OGBV normalises misogyny and promotes a culture of patriarchal violence, involving rape culture, victim-survivor blaming and trivialising sexual assault.\(^7^9\) The normalisation of OGBV reinforces a ‘silencing’ of women and LGBTQ+ people, whereby the victim-survivor is discouraged from participating in public life. OGBV thereby exacerbates gender inequality that limits women and LGBTQ+ community from exercising their freedoms and human rights. There is also an intergenerational impact as OGBV deters and impedes young women and girls and LGBTQ+ people entering professions such as politics and journalism, due to fear of similar abuse, which, in turn, increases the gender digital divide.\(^8^0\)

- **Threats to private and public safety:** Reiterating that OGBV occurs within a continuum, misogynistic behaviour that starts in the online space may lead to the perpetration of offline violence – both in private and public spheres.\(^8^1\) For example, a 2023 US Secret Service report details the public security threat posed by individuals who perpetrate acts of targeted violence, with attackers engaging in domestic violence, misogynistic behaviours, or both prior to an attack. It notes that men who have committed misogynistic violence (typically mass shooting and stabbings) have histories of concerning and threatening online communications, as well as other risk factors (such as a history of being bullied, financial instability, and interpersonal difficulties).\(^8^2\)

- **Threats to democracy:** At a societal and global level, anti-democratic forces – both foreign state and non-state malign actors – exploit online spaces to attack women and LGBTQ+ people in public life.\(^8^3\) A report by #ShePersisted notes that gendered disinformation can serve as an early-warning system for “both backsliding on women’s rights and the erosion of democratic principles and institutions.”\(^8^4\) A global study conducted by ICFJ notes an “alarming trend” of the role played by political actors, including politicians, government officials, political party representatives, party members, political operatives, and extremists on the political fringe, as “instigators and primary perpetrators of online violence against women journalists.”\(^8^5\) A global report on gendered disinformation by the U.S., Canada, the European External Action Service, Germany, Slovakia, and the UK further emphasises that foreign state actors like Russia and the People’s Republic of China (PRC) strategically target women and people with intersecting identities to dissuade individuals and identity-based groups from exercising their rights. The report further asserts that identity-based disinformation undermines the “ability to access impartial, fact-based information, and it negatively impacts the make-up of democratic representation.”\(^8^6\)
 Assessment of platform policies and enforcement

A key factor in responding to OGBV is the development and effective enforcement of comprehensive community guidelines or standards, which outline what is and what is not allowed on a platform. These are generally contracts of adhesion, presented to users on a take-it-or-leave-it basis, and include a set of policies that are frequently updated by the platforms.

While most platforms generally account for some forms of OGBV in their hate speech or harassment and abuse policies (including the protected characteristics of users), this section outlines some of the gaps in both policy and enforcement. In terms of the former, none of the platforms explicitly address gendered or sexualised mis- and disinformation. Yet, such content often comes in the form of coded and veiled language, context-based visual and textual memes, or use tactics of intentionally obscuring certain words. Furthermore, the Oversight Board, which reviews content decisions made by Meta, recently overturned Meta’s decision to keep online a Facebook post that mocks a victim-survivor of GBV. Specifically, the Board found that the post violated Meta’s Bullying and Harassment policy as it mocked the serious physical injury of the woman depicted. The Board explained, however, that “this post would not have violated Meta’s rules if the woman depicted was not identifiable, or if the same caption had accompanied a picture of a fictional character,” indicating a gap in policy that seems to allow content that normalises GBV.

Beyond gaps in policy formulation, ISD research identified patchy enforcement of existing policies. On X (formerly Twitter), which prohibits “targeting others with repeated slurs, tropes or other content that intends to degrade or reinforce negative or harmful stereotypes about a protected category,” ISD research conducted in the US context found multiple instances where this type of content was not moderated, including tweets containing sexist tropes against the actress Amber Heard as well as general attacks on women’s appearances. The same research also identified openly derogatory terms such as “whore”, “cunt” or “bitch” in the comment section of YouTube, which would often occur not only under videos that seemed to invite hateful comments but also under inconspicuous videos (e.g., such comments were found beneath both “Andrew Tate Destroys Modern Women” and “House Speaker Nancy Pelosi holds her final weekly press conference”). On TikTok, ISD found that misogynistic content is still openly posted and promoted by users, including videos by accounts promoting Men Going Their Own Way (MGTOW), a sub-movement of the Manosphere, that among other things, belittles and dehumanises single mothers and their role in society.

The platform with the most lax (or lacking) policies and enforcement (i.e., where ISD found the most misogynistic content) is Telegram. Notably, ‘private’ channels (which require an invite by the owner or an invite link to join, but in practice are often easily joined) are not covered by the terms of service. As of the time of writing, the service has no policies addressing hate speech, nor does it prohibit doxing, despite having been criticised for hosting “an epidemic of politically motivated doxing, allowing dangerous content to proliferate, leading to intimidation, violence, and deaths.” Researchers noted that while Telegram was designed as a messenger service, it has become a hybrid between a messenger service and a social media platform as messages in public channels can reach hundreds of thousands (or even million) views, with some channels enabling commenting on posts or reacting to them via emojis. Moreover, ISD found that sharing links to other audio-visual platforms, such as video and livestreaming websites, is particularly popular among right-wing extremists and conspiracy theorists on Telegram. This reiterates the challenge of cross-platform dissemination of content, coordination and activities, including the relevance of accounting for the risks posed by smaller and ‘alternative’ platforms.

Finally, a well-observed problem is the lack of sufficient moderation resources for content published diverse local contexts, including in non-Western regions or non-English languages. For example, a report by the Slovak Council for Media Services and Reset reviewed the role of platforms in the case of a shooting outside an LGBTQ+ bar in Bratislava in 2022, finding that Facebook had sent reported content to a hired third-party fact-checker to perform the review. However, there was only one Facebook-contracted fact-checker for all of Slovakia, showcasing how limited resources pose obstacles for the rapid and efficient review of content violating the community guidelines.
Platform design and systems: Risks of reproducing and amplifying OGBV

Studying the impact of platform design and systems on exacerbating the risks of OGBV and the dissemination of misogynistic content remains a challenge given that users experience highly personalised online interfaces and spaces. This section reviews how platform design and systems risk reproducing and amplifying OGBV, assessing user interface design and artificial intelligence (AI)-based systems, including the use of algorithms and machine learning (ML) models for personalised feeds and (search) recommendations.

User interface design: A relatively well-studied phenomenon showcasing how design can be harmful to users are deceptive patterns, also known as “dark patterns”. There are many forms of deceptive design patterns, but essentially, they are “choices that can [unintentionally or intentionally] influence or trick users into making unintended decisions.”

In 2021, interdisciplinary research conducted by Caroline Sinders, Vandini Shukla and Elyse Voegeli surveyed journalists’ interactions and relationships with platforms’ user interfaces, given their status as a user group that faces a range of harassment and harm online, including gendered or racial slurs, doxing, and rape threats. Their research emphasises that “technology is a planned space, and users can only conduct specific actions that are designed and allowed by the software, application or platform they are using.” The design of online spaces is thereby closely interlinked with the experience of potential harms on platforms. Deceptive design choices can negatively impact users’ privacy and safety, for example:

- Settings that default to the least privacy friendly option;
- Rewards and restrictions if users decline or opt out of settings, such as loss of functionalities;
- Forced action to complete the settings review at a time determined by the platform, pressuring users without a clear option to postpone the process;
- An illusion of control as the platform provides users with granular choices that ultimately discourage them from changing or taking control of their settings.

Deceptive design highlights the importance of applying a Safety by Design approach, which encourages platforms to build safety into the design, development, and deployment of their features, rather than retrofitting safety solutions after harms have occurred.

The survey also indicates a lack of victim-survivor-centred reporting mechanisms and communication of community guidelines. Surveyed journalists experienced harassments in peaks, with patterns of harassment instead of stand-alone instances. However, they were only able to report individual instances rather than multiple instances in bulk. Journalists also expressed “frustration and confusion over how platforms responded to harassing content.” It is important to consider that limited user agency in the face of this type of harassment may replicate a loss of power, control, and rights, which is an experience shared by many victims-survivors of OGBV. The coordination of harassment campaigns across platforms and the lack of interoperable reporting mechanisms further weaken user agency and risk mitigation.

AI-based systems: Personalised feeds and other AI-based services such as search recommendations shape the user experience, creating specific risks of OGBV. Algorithmic systems make automated decisions that score and rank content and suggestions for who to connect with (or what pages or groups to follow) based on signals, including users’ historical behaviour (such as viewing history) and predictions derived from past behaviour of similar users (using techniques such as collaborative filtering). ML models use predicted probability of engagement (the probability of users liking, sharing, viewing, etc. content) to optimise the order in which content is ranked and displayed on user feeds. In short, the goal of engagement-based ranking is to maximise whatever engagement goal (metric) a company has set, often at the level of individual user indicators (for example, the time users spend on a platform).

While studying the impact of algorithmic feeds on discourses and user behaviour remains challenging, there is evidence pointing toward an engagement problem, which describes the tendency to engage more with content that is low-quality (such as clickbait...
headlines) or nears a “cut-off point” of what is allowed under the community guidelines (borderline content). Therefore, engagement-ranked feeds often risk creating a “gravitation” towards borderline content, which also increases the risks of recommending misogynistic content.

For example, ISD research suggests that platforms give greater visibility to abusive hashtags over non-abusive hashtags. On Instagram, the transphobic hashtag #rachellevineisaman (25 posts) was ranked third among recommendations, ahead of non-abusive hashtags featured in more posts (e.g., #rachellevinephotography with 183 posts). Moreover, research finds that Tradwives are able to “adapt their content” to exploit algorithmic feeds, using self-branding strategies, presenting the “#tradlife” and sharing “homemaking, cleaning, and beauty content” to engage audiences, while promoting antifeminist and anti-LGBTQ+ belief systems. They further capitalise on engagement-based ranking by commodifying far-right ideology through advertising, brand collaboration, or promotions.

Studying AI-based systems in general, scholars suggest that algorithms reflect and exacerbate gender norms already present in society. Patterns of gender bias and discrimination have been detected in algorithms used for hiring decisions, criminal sentencing, and health-care allocation, among others. Researchers from New York University (NYU) demonstrated that societal levels of inequality are evident in search algorithms, noting a “cycle of bias propagation between society and AI.” This reiterates the need to address the challenges of AI for society, including how it risks reproducing and amplifying OGBV, in a multistakeholder effort that involves interdisciplinary research and development of offline and online responses.

As outlined, platforms employ algorithms and ML models to make predictions about user engagement, using large amounts of user data. However, automated inferences risk biased outputs, including gender stereotyping, given that datasets often contain racial and gender biases. For example, researchers tested the accuracy of X’s (formerly Twitter) inferences of users’ gender identities, finding that the LGBTQ+ community and straight women were more often misgendered than straight men. Researchers emphasise that misgendering users, “beyond echoing deeply rooted stereotypes, can lead to privacy and discrimination issues.”

Rebekah Tromble, Director of George Washington University’s Institute for Data, Democracy and Politics, describes the problem this way: “how we consume social media content is an inherent human construct. And if there are problems with how this consumption happens, it’s down to concerted decisions from social media executives and engineers — and not some natural phenomenon that is out of anyone’s hands.” How platforms design and evaluate their algorithms directly impacts user experiences, especially as they risk amplifying the dissemination of misogynistic content.
Response measures: Risk assessment and mitigation of OGBV

Based on the review of trends in OGBV and platform policies, design and systems, the following sections evaluate and propose how platforms can work toward better assessing and mitigating OGBV and its connections with violent extremism. Recognising the need for multistakeholder approaches and solutions, this section considers proposed measures at the multilateral and governmental level, as well as by industry, civil society, and academia.

Enable API access to platform data and develop standardised transparency reporting

The monitoring, measurement, and transparent reporting of OGBV by platforms are prerequisites to understand and explain the nature, scale, and scope of the phenomenon. Additionally, API access to publicly accessible data should support public interest research and enable evidence-based decision-making.

The Global Partnership together with UN Women, the WHO, UNFPA and UNICEF initiated efforts toward enabling the production of accurate, reliable and comparable data and knowledge around OGBV. In 2023, the UN Women-WHO Joint programme on Violence against Women (VAW) data published a paper on ‘Taking stock of evidence and data collection’, scoping methodologies and recommendations on the approaches to collecting data on Technology-Facilitated Violence against Women (TFVAW). The paper highlights existing methodologies as well as methodological, ethical and socio-political challenges. These include the lack of “overall problematisation and awareness” around TFVAW due to a lack of data and dissemination of research findings, and a bias of data towards the Global North, neglecting the differentiated impacts across diverse and different contexts. The paper highlights the need for a shared operational definition and methodology for monitoring, measuring and analysing TFVAW. It further notes the importance of incorporating social media data and the need to consider a “diversity of methodologies” to allow for different data sources.

In this context, the collection and analysis of platform data should address data gaps for the purpose of evidencing the tactics and forms of OGBV as well as the connections between misogyny and different extremist ideologies. For example, API access could allow cross-ideological analysis of different violent extremist and terrorist actors, including a comparative analysis of the respective gender dimensions. Vetted researchers from different disciplines such as Computational Linguistics, Critical Terrorism Studies, or Critical Studies on Men and Masculinities should have meaningful API access to systematically collect and analyse data. Regulatory frameworks already address the need for such data access. Notably, Article 40 of the European Union (EU)’s Digital Services Act (DSA) requires that access to data “publicly accessible in their online interface” should be made available, where possible, in real-time to researchers, including those affiliated to not-for-profit bodies, organisations and associations. In parallel, company signatories of the 2022 Strengthened Code of Practice on Disinformation committed to voluntary standards that will serve as co-regulatory measures for the DSA. The Code includes the commitment to “continuous, real-time or near real-time, searchable stable access to non-personal data and anonymised, aggregated, or manifestly-made public data for research purposes on Disinformation through automated means such as APIs.” Available platform data should be compared and triangulated with data from other sources such as administrative data, statistics, or surveys to ensure a comprehensive mapping of the phenomena.

In addition, platforms should develop standardised transparency reporting to include gender-aggregated data to allow external researchers to scrutinise and track the enforcement of policies, especially considering violations of hate speech and TVEC policies. For example, enforcement reports should include aggregated data on the prevalence of and user engagement with content (including but limited to posts, comments, and profiles) detected as gender-based hate speech, the proportion of image-based content that violated these policies, as well as data on how user reporting was addressed (e.g., what specific actions were taken). Enforcement reports should also account for hate speech directed towards other protected groups to measure intersecting identity-based hate and support intersectional analysis of the motivations driving OGBV.

Platforms should work with GBV and feminist advocates, scholars, and victims-survivors of OGBV when developing methodologies of transparency.
reports (including content categories and metrics), or when conducting internal research (such as user surveys about experiences of OGBV). Considering the lack of a universally agreed definition of OGBV and the need for more consistency of transparency reporting (and thereby comparability of platform actions), a cross-sector effort could also contribute to and participate in the ongoing work by UN Women\(^{124}\) to develop terminology and a statistical framework for TFGBV and the UN Special Rapporteur on Freedom of Expression to develop a common definition for gendered disinformation.\(^{125}\)

**Apply a victim-survivor-centred Safety and Privacy by Design approach**

Scholarly and policy debates recognise the need to take steps to provide short-term relief and mitigate the risks of OGBV. While user interface design can undermine user agency and safety (e.g., recalling dark patterns outlined above), it can equally enable users to mitigate risks of misogyny.

Reviewing and evaluating immediate responses, research by Sinders, Shukla and Voegeli (2021)\(^{126}\) and PEN America\(^{127}\) emphasises the need for platforms to implement improved user tools. Their recommendations propose proactive measures that enable users to reduce risks and exposure to OGBV, reactive measures that facilitate more effective immediate responses when users are faced by OGBV, and accountability measures to aim to deter abusive behaviour and discourage perpetrators from exploiting platforms for networking and coordinated harassment.

A Safety and Privacy by Design approach centres user agency in the development and design of platform products and services. The following measures are not exhaustive, and more research will be needed to evaluate their effectiveness.

**Proactive measures may include:**

- Content moderation tools such as “shields” that enable users to proactively filter abusive content (across feeds, threads, comments, replies, direct messages, etc.) and quarantine it in a dashboard, where they can review and address it with trusted allies;
- Robust, intuitive, user-friendly features that allow a fine-tuning of privacy and security settings, including “visibility snapshots” that show, in real time, how adjusting settings affects reach;
- Structures that allow users to assemble rapid response teams of trusted allies, including the delegation of account access.

**Reactive measures may include:**

- Emergency hotlines that users can use to receive trauma-informed support in real time;
- Documentation features that allow users to record evidence of OGBV quickly and easily (for example, instantly capturing screenshots, hyperlinks, and other publicly available data), which should be made interoperable to allow cross-platform evidencing;
- Improved and standardised features to block contacts, mute content, and restrict or hide content;
- Improved reporting mechanisms, including bulk reporting in recognition of coordinated nature of harassment campaigns, as well as circular reporting that allows for a report to be reopened and edited, and across platforms.

**Accountability measures may include:**

- A transparent system of “escalating penalties” for abusive behaviour, including warnings, strikes, nudges, temporary functionality limitations, suspensions, content takedowns, and account bans. In terms of account bans and de-platforming, research has noted that “removing perpetrators may not get at the root of the problem of accountability,” while emphasising that “lock-down mechanisms” should preserve metadata and account information for evidence-gathering and accountability-related purposes;\(^{128}\)
- Testing “proactive nudges” that aim to encourage users to revise abusive content before they post it (as well as research measuring the efficacy of nudges);
- Sufficiently resourced appeal processes to ensure the clear and time-sensitive review of appeals.
Legislative frameworks have identified the need for user agency and empowerment, and started requiring platforms to be more accountable and transparent about their design and policies. For example, Australia’s Online Safety Act 2021 refers to Basic Online Safety Expectations, which require platforms to put in place clear and readily identifiable mechanisms that enable users to report and make complaints about content as well as terms of use, policies, and procedures to deal with complaints and reports. The expectations also require that platforms keep records of user reports and complaints for five years.

**Harassment Manager developed by Google’s Jigsaw**

Harassment Manager is an “open source codebase for a web application that allows users to document and manage abuse targeted at them on social media,” starting with X (formerly Twitter), who partnered on the project. The tool intends to help users “identify and document harmful posts, mute or block perpetrators of harassment and hide harassing replies to their own tweets.” Users can review tweets based on hashtag, username, keyword or date, leveraging the Perspective API to detect comments that are most likely to be toxic (further discussed below). The Harassment Manager code is available on Github, open sourced for developers and non-governmental organisations to build and adapt for free. This tool should be tested and scrutinised by GBV and feminist advocates and experts as well as victim-survivors to inform the further improvement and development.

**Enhance cross-platform cooperation and information sharing**

Platforms should recognise that what occurs on other platforms may make its way to their own service (and vice versa). This is not only true of TVEC, but also of the actors and tactics of OGBV. Online harassment campaigns targeting an individual may be coordinated on one platform, with the content or URLs to this content cross-posted to other platforms, where the targeted users may or may not have accounts. As stated earlier, harassment or the coordination of harassment often also involves smaller and ‘alternative’ platforms.

Platforms should develop and operate exchange channels between relevant teams, including safety and content moderation teams, to proactively share information about cross-platform harassment (such as perpetrators using multiple accounts), including, where relevant, user reports of multi-platform harassment. Platforms should also develop interoperable reporting mechanisms for users to enable user agency and efficient response. Exchange channels may facilitate faster action, for example, when a prominent actor is identified to be linked to repeated harmful behaviour, such as violating community guidelines across platforms. Such efforts are important for understanding the scope and nature of OGBV, but also to coordinate mitigation actions by platforms and other stakeholders, including governments, civil society, or even law enforcement, if relevant.

Already existing cross-platform efforts and crisis protocols such as the Global Internet Forum to Counter Terrorism’s Content Incident Protocol and the Christchurch Call Crisis Response Protocol should consider how OGBV is relevant to their scope and mandates, and how to strengthen mechanisms appropriately, recognising misogyny as a radicalisation vector for violent extremism. Relevant voluntary commitments or co-regulatory frameworks could also be reviewed. For example, the EU’s 2022 Strengthened Code of Practice on Disinformation includes the commitment to “operate channels of exchange between their relevant teams in order to proactively share information about cross-platform influence operations, foreign interference in information space and relevant incidents that emerge on their respective services, with the aim of preventing dissemination and resurgence on other services.” Committed channels could extend to incidents of gendered or sexualised harassment campaigns. Such an effort could also be seen as beneficial to compliance with the EU’s DSA under which platforms are required to assess and mitigate systemic risks related to OGBV.

**Review and update content moderation policies, processes, and systems**

Platforms should assess and mitigate how patriarchal gender norms factor into and are reproduced by their moderation policies and practices. A comprehensive approach to community guidelines and moderation that addresses OGBV, applying a gender lens and a victim-survivor-centred approach, should sensitise policy formulation and enforcement to the continuum of
OGBV as well as the links between misogyny and violent extremism, including the use of dehumanising language based protected characteristics of persons.

Relevant platform teams should consider conducting interviews and focus groups with victims-survivors to inform policy and enforcement processes. Gaps in moderators’ lack of understanding of local languages and regional contexts need to be addressed by involving diverse population groups. Civil society has also suggested that platforms provide support, including trauma support, to local organisations who review hate speech policies and develop local lexicons of misogynistic words and phrases.

An example of how AI-based systems could be used to support content moderation is Perspective API, developed by Google’s Jigsaw, which uses ML models to identify abusive comments online. Perspective API predicts the perceived impact a comment may have on a conversation by evaluating (scoring) that comment across a range of emotional concepts (attributes). Currently, Perspective API may provide scores for attributes defined as “Toxicity”, “Severe Toxicity”, “Insult”, “Profanity”, “Identity attack”, “Threat”, and “Sexually explicit”. The tool intends to help moderators to quickly prioritise and review comments that have been reported, to give feedback to commentators, and for users to control which comments they see. It proposes an encouraging outlook for applying AI-based systems to improve content moderation and to make the online environment safer. Noting that the development and application of such tools is in the early phases, researchers have tested Perspective API to measure levels of toxicity of tweets from prominent drag queens in the US. The research suggests that Perspective considered a significant number of drag queen accounts to have higher levels of toxicity than accounts of white nationalists. Thereby, it was not able to consider the social context when measuring toxicity levels, for example, it did not recognise cases in which words, that might conventionally be seen as offensive, conveyed different meanings in LGBTQ+ speech. This suggests the need for continued and increased multistakeholder collaborations to build on and advance industry tools such as Perspective API.

Both AI-based and human content moderation require comprehensive and regular updates of policies, including trauma-informed processes, to address the nuanced forms of OGBV and prevent counter-productive outcomes. These efforts require a genuine will to improve systems and should not be negatively influenced by company metrics that prioritise engagement. Instead, platform actions should prioritise the principle of ‘do no harm’, mitigating the exposure to risks.

Audit and mitigate misogyny in AI-based systems

Legislation such as the EU’s DSA as well as proposed (non-binding) guidance such as the Violence Against Women and Girls (VAWG) Code of Practice proposed by advocates in the context of the UK’s Online Safety Bill (OSB) intend to assess and mitigate the gendered impact of platform design, including their algorithmic feeds.

Articles 34 and 35 of the DSA specifically call on platforms to assess and mitigate the systemic risks posed by their services, including any negative effects in relation to OGBV. Thereby, risk assessments should include the “design of their recommender systems and any other relevant algorithmic system.” The VAWG Code emphasises the Safety by Design principle to ensure that “algorithms used on the service do not cause foreseeable harm through promoting hateful content, for example by rewarding misogynistic influencers with greater reach.” The Code argues that “preventative measures must consider the role of algorithmic product decisions,” reiterating that the decision-making processes around their development and deployment must be scrutinised.

Algorithmic accountability and auditing should take a victim-survivor-centred approach and conduct safety testing, and apply gender analysis and intersectional perspectives, specifically testing how individual users experience intersecting forms of identity-based hate and violence.

Auditing and evaluating the impact of algorithms remains a challenge, even with direct access to proprietary code, given that algorithmic feeds are personalised and rely on many factors including users’ historical data. Moreover, independent auditors need to use a counterfactual scenario to compare the algorithmic feed, for example, when conducting randomised controlled experiments. For example, a recent study assigned a sample of consenting users to reverse-chronologically-ordered feeds to assess the impact of algorithmic feeds, including
how they encourage partisan stereotyping or influence negative attitudes about outgroups. In the context of OGBV, studying the role of algorithmic recommendations in the Manosphere may help with developing evidence-based interventions in online radicalisation pathways, given that it is likely that users who enter the Manosphere may have less intense, less extreme beliefs and slowly form new connections and become further embedded within the inner community.

Christchurch Call Initiative on Algorithmic Outcomes

The Christchurch Call Initiative on Algorithmic Outcomes, led by New Zealand, the US, X (formerly Twitter) and Microsoft, seeks to develop software tools to facilitate independent research on the impact of user interactions with algorithmic systems. Working with OpenMined, DailyMotion and LinkedIn, a new software infrastructure will integrate privacy enhancing technologies to allow external researchers and data scientists to remotely study algorithms distributed across multiple secure sites. Such effort is crucial to enable independent research on the impact of algorithmic feeds. The independent auditing of algorithms and ML models via such software infrastructure should focus on understanding and testing the role of algorithmic pathways, including, where possible, across platforms. The development of systems for remote researcher queries will need appropriate governance and ethics frameworks as well as processes for research prioritisation.

In terms of potential mitigation of biased AI systems, scholars have suggested that ML models can be developed so that they do not produce discriminatory patterns such as gender stereotypes. The idea would be not to limit the data input (i.e., remove any data related to gender), but to prevent algorithms from yielding gender-based patterns, since not using gender data may still allow for predicting gender and result in discrimination by proxy. For example, risk mitigation could involve interventions for bias reduction, including debiasing an algorithm’s training set. Transparency and inclusivity, by incorporating intersectional feminist knowledge, will be critical for algorithmic auditing.
Conclusion

Normalising misogynistic violence, the harassment and intimidation of women and the LGBTQ+ community, and upholding patriarchal gender norms, are all situated in “larger patterns of systemic violence made to control, demean, and significantly limit the autonomy of the person targeted.” Online manifestations of GBV impede the safety, freedom of expression, and participation in public life of women and LGBTQ+ people.

In this context, the paper has emphasised the continuum of violence within which misogyny can serve as an ideological link across different forms of violent extremism. A recognition of the systemic issue of patriarchal norms in society and the risk of misogyny as an ideology that can be a gateway to radicalisation should be reflected in a systemic response by platforms, governments, and civil society.

Beyond immediate action to enable victim-survivor-centred user agency, platforms should assess both individual and societal level harm caused by OGBV, especially recognising the need to consider the relationship between misogyny and TVEC in their community guidelines and risk assessments. Evidence gaps in the research on violent extremism and misogyny reiterate the need to further study these complex phenomena, including by means of strengthening data access and transparency reporting. Evidence-based decision-making and interventions are central to avoid any potential negative consequences for achieving gender equality and safeguarding freedom of expression.

Platforms should develop inclusive community guidelines and sufficiently invest in clear and consistent enforcement. Platforms should also assess and mitigate any risks stemming from the functioning of their systems, which include algorithmic recommender systems. Notably, product teams should assess how algorithmic design (and unintended consequences) reflect and reproduce patriarchal gender norms that risk amplifying misogynistic content. A Safety by Design approach should ensure inclusive design and safety testing that incorporates intersectional perspectives before launching new services and features. Finally, meaningful access to platform data via APIs for vetted researchers remains fundamental to gathering and understanding evidence, making sense of research findings, and holding platforms accountable. This paper and its recommendations should be understood as complementing whole-of-society and whole-of-government actions as platforms can play a crucial role in supporting such efforts.
Endnotes


Recommended Measures for Platforms to Assess and Mitigate Online Gender-Based Violence 25


26 Ibid.


35 Ibid.


87 ISD research reviewed the policies of X (formerly Twitter), Meta’s Facebook and Instagram, YouTube, TikTok and Telegram.


It is important to consider that the type of data deemed “publicly accessible” can change. Thereby, tracking any changes should build on a shared taxonomy of what data is “publicly accessible.” Public interest research commonly refers to research with the explicit aim to develop society’s collective knowledge. Regulatory precedent suggests that public interest research must be independent of commercial interests and reveal the source of its funding. Public interest researchers are not necessarily linked to academic institutions and can also include researchers affiliated to non-profit or media organisations.

Public interest research commonly refers to research with the explicit aim to develop society’s collective knowledge. Regulatory precedent suggests that public interest research must be independent of commercial interests and reveal the source of its funding. Public interest researchers are not necessarily linked to academic institutions and can also include researchers affiliated to non-profit or media organisations.

Methodologies include survey data generated through mixed-methods, including by means of machine learning to collect and analyse social media data.


Context

- President Emmanuel Macron of France and Prime Minister Jacinda Ardern of New Zealand launched the Christchurch Call on 15 May 2019, along with Leaders from governments, online service providers, and civil society.
- The Call is a response to the online elements of the livestreamed attack on two mosques in Christchurch, New Zealand, in which 51 people were murdered.
- The Call is a multi-stakeholder initiative, and contains 25 commitments aimed at eliminating terrorist and violent extremist content Online, while preserving human rights and a free, open, secure internet.
- At the 2022 Summit, Leaders from across the Call Community identified a need to engage with children and young people.
- Youth are vulnerable to the impacts of radicalisation online, and have valuable expertise to offer in countering radicalisation and TVEC.
- The Call Community has developed a youth engagement plan and will soon convene a Christchurch Call Youth Reference Group. This meeting provides an opportunity to directly discuss our work with young people and to seek their views.

Purpose and Objectives

- To hear directly from young people about how we can ensure our engagements with youth are genuine and effective.
- To improve our understanding about how to keep young people actively engaged in these conversations.
- To better understand which TVEC-related issues are priorities for young people, to subsequently inform Youth Reference Group discussions.
- To improve our understanding of how children and young people are using the internet, what platforms they are using, and what policymakers might be missing in their work.
- To receive preliminary advice on what young people are seeing and engaging with online, how they are using online environments, and how we might counter youth radicalisation.

Logistical Details

- **Format:** The first 40 minutes will feature small group discussions around two key themes: youth engagement and young peoples’ views on terrorist and violent
extremist content online. In the last 20 minutes, participants will be invited to share their views with senior representatives from the Christchurch Call Community. The discussion will be held in English.

- **Ground rules:** The Christchurch Call Community is committed to remaining a safe space for open dialogue. As such, we use Chatham House Rules in our meetings, meaning that you are welcome to share the content of this discussion, but may not share information about who has made a contribution.

- **Pre-meeting process:** Registrants will be invited to join a Teams meeting prior to the session in order to build relationships, to ask questions about the event, and to identify priority issues for the session.

- **Post-meeting process:** This conversation will help us shape the focus and engagement strategy for the Christchurch Call’s Youth Reference Group. Advice gathered in the meeting will be shared with Leaders in the Call Community and with the Community more broadly. A summary of advice will also be shared on the Christchurch Call website.

- **Registration:** Please email plenary@christchurchcall.com to register for this event. Please include your full name, email address, organisation (if applicable) and gender. Please return these details by **October 25 2023.**
Co-Chair Statement
Christchurch Call Leaders’ Summit
NEW YORK, 20 SEPTEMBER 2022

Joint Statement by Prime Minister Rt Hon Jacinda Ardern and His Excellency President Emmanuel Macron on the occasion of the 2022 Christchurch Call Leaders’ Meeting
Today, Leaders of the Christchurch Call Community – the Heads of State and Government, Leaders from Online Service Providers and Civil Society have convened to address our continued efforts to combat terrorist and violent extremist content online.

Just over three years ago at the Elysée Palace in Paris, Leaders from the technology industry and Government were joined with a karanga – a call to attention from the indigenous culture of Aotearoa New Zealand, remembering not only the lives lost in Christchurch but also countless more lives and communities around the world impacted by the weaponization of the internet and online services.

The Christchurch Call made significant progress. Together, among other important actions, we have:

- Enforced new policies and set up systems to combat terrorist and violent extremist content;
- Supported establishment of the Global Internet Forum to Counter Terrorism as an independent NGO to create integrated solutions and share information and expertise;
- Developed incident protocols that increased the speed and effectiveness of our response to livestreamed terrorist attacks and similar incidents; and
- Built a multistakeholder community that spans every continent, dedicated to advancing our work together.

Since our last Summit, we have added new industry supporters and partner organisations to our ranks including Roblox, Zoom, Mega, Clubhouse, the Global Community Engagement and Resilience Fund, and Tech Against Terrorism.

The karanga, or call, was not just to those who have passed, but also to future generations. The future of the internet and the future of our free, open, societies are intertwined. As we look ahead our legacy must be a free, open, secure and interconnected global internet as a force for good, a place where human rights are promoted and upheld and where technology contributes to social mobility and empowerment for all.

Today, we considered how to take forward our work in three broad areas.

1. Incident Response and the Proliferation of Terrorist and Violent Extremist Content

The Christchurch Call has galvanised significant action, meaning users are far less likely to encounter terrorist and violent extremist content online. We have also built systems to coordinate and respond during a terrorist attack, to make it more difficult for attackers to exploit online services. There are, however, many online service providers that remain outside the Call. Some are small. Others cannot or will not meet the Call Commitments. The threat landscape continues to evolve. We have seen concerted efforts to exploit gaps and vulnerabilities and find novel ways to defeat the detection and moderation efforts of Call supporters. We therefore need to extend our reach and to harness the commitments, capabilities, skills, and networks that exist across all sectors.

Today the Call’s Leaders endorsed the following key actions:

- Continue reaching out to a diverse set of online service providers both large and small, bringing them into the Call and into our systems for crisis response.
• Ensure shared industry tools remain at the leading edge of technology and are effective in responding to adversarial tactics and shifts.

• Start Community work on options and advice for Call Supporters to address the role of unmoderated and ‘alt-tech’ services in disseminating terrorist and violent extremist content, day-to-day and in crises, while protecting and respecting human rights and maintaining a free, open, and secure internet.

• Continue working with all sectors to test and refine our crisis response protocols, to improve information-sharing and coordination, to strengthen human rights protections, and to increase transparency. Ensure an active role for researchers, civil society, and community groups in crisis response.

2. Algorithms, social drivers, impacts, and interventions

Our Mission is to deal with the online impacts of terrorism and violent extremism and to address its drivers including underlying social factors and the impact of online user journeys on radicalisation to violence. Enhancing our understanding of the impacts of algorithmic systems – including those which curate and recommend the content we see, and those that identify and remove violating content – is a crucial element in making the internet safer. There are many challenges that hinder the effective investigation of algorithmic systems including commercial sensitivity, privacy of user information, and the technical challenge of accessing systems and data securely. Insights derived from improved understanding of these processes will enable us to develop more effective interventions.

The Call’s Leaders endorsed the following key actions:

• Develop shared solutions to studying algorithmic impacts within the Christchurch Call

• Drive forward discussions on targeted research pilots that respond to questions raised by our Community, including on how content recommendations might be exploited by terrorists, deepening the evidence base on the links between online misogyny as a vector for violent extremism, and the possible unintended consequences of human / machine learning AI interactions

• Promote a wide range of tools, including third party standards and quality metrics, and features that promote user choice and responsibility to help disrupt violent extremist and terrorist content.

• Increase outreach to organisations representing communities impacted by terrorism and violent extremism online, including women, LGBTQIA+ communities, youth, and intersectional communities, to embed diverse voices in our understanding of the impact of algorithms, radicalisation, and the design and implementation of effective interventions.

• Recognising that there are demonstrated linkages between targeted violence and online hate-based movements and ideologies, including, for example, online misogyny, gender-based hatred, and other hate-based, discriminatory prejudices, deepen the evidence base on the links between these factors as potential vectors for terrorism and violent extremism.

3. Future-proofing the Christchurch Call

As we continue to innovate and the functionality of online environments grows, we want today’s young people to enjoy the benefits of a global internet without having to confront violent extremist and
terrorist content or threats. We are determined that the Call will play a role in building a positive future online, and that our Call Community continues to thrive and contribute positively to it.

The Call’s Leaders endorsed the following key actions:

- Launch a new stream of work to understand how we can support the adoption of new technologies while promoting safety and securing against terrorist and violent extremist content.
- Increase our understanding of the challenges presented by new technologies, to develop strategies to address those challenges and prepare the Call Community to counter new types of terrorist and violent extremist content.
- Engage directly and thoughtfully with children and young people about our new workplan and actions under the Christchurch Call and reflect their voice in our work.
- Prioritise outreach to online service providers that serve a younger audience as we build our multistakeholder community, with a focus on those at the forefront of emerging technologies such as immersive gaming platforms.
- Deliver a platform to help our Community work more effectively together on common projects.
- Build on the success of the Call as a global multistakeholder Leaders initiative by pushing for globally aligned approaches and solutions.

The Christchurch Call has developed a unique model for coordinating action, bringing together affected communities, civil society, and technical experts, alongside international organisations, industry and government. By applying the distinct capabilities of each sector and forging a community with shared purpose and ambition, we are delivering results.

The success of the Call is now well recognised, and some stakeholders have expressed interest in seeking to understand how the Call might work on related issues. We agreed the Call should remain focused on the scope of its Commitments building on our success and sustaining this work into the future.

Ahead of the Paris Peace Forum, New Zealand, France, and other Call community members with an interest will consider some of these related issues – including disinformation, harassment, abuse, and hatred online, and issues affecting youth – to understand how we might apply what we have learned working on the Christchurch Call, where we can support and engage on related initiatives such as Tech for Democracy, the Summit for Democracy, the Freedom Online Coalition, the Declaration for the Future of the Internet, the Aqaba Process, the Global Partnership for Action on Gender Based Online Harassment and Abuse, the Global Partnership on Artificial Intelligence, and the International Call to Stand up for Children’s Rights Online, and where there is multistakeholder interest in new work programmes separate to the Call.

We look forward to re-convening Christchurch Call Leaders in 2023.