Algorithms & Positive Interventions Workstream

Objective of the Workstream

1. To provide impetus and momentum to work to fulfil the Christchurch Call commitments undertaken by Online Service Providers, and Governments, including supporting work already taking place (e.g. in fora such as GPAI and GIFCT's CAPPI working group), as well as identify any gaps and areas where further work might be needed.

The workstream recognises that preparing this document in a short space of time is a challenge and will require us to focus on the areas likely to be most relevant and important for Tech, CSO and Political Leaders to take up.

Relevant Christchurch Call Commitments made by Governments and Online Service Providers (Paris, 2019)

- Governments to **Consider appropriate action** to prevent the use of online services to disseminate terrorist and violent extremist content (TVEC), including through collaborative actions, such as: […] development of industry standards or voluntary frameworks; regulatory or policy measures consistent with a free, open and secure internet and international human rights law.

- Online Service Providers to **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.

- **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.

- Governments and Online Service Providers to **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.

- **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.

Key Issues/Challenges Identified by the Workstream

Principles and landscape

1. Freedom of speech is an important component of the Call commitments. However, there are concerns that products and design features may facilitate user engagement or lead

users towards terrorist or violent extremist content. Some participants reflected that free speech doesn't equate to freedom of reach while others see this as a more grey area for instance in the context of a marginalised community's ability to be widely heard or suppressed. Some Call Community members pointed to the importance of acknowledging the context of power dynamics and imbalances.

2. A particular challenge in looking at the problem, is the extent to which algorithms may suppress legitimate conversations on controversial issues in seeking to address terrorist content online. One way of addressing this challenge is to focus on the user experience, and particularly any processes/algorithms that may drive a user towards and / or amplify TVEC (in accordance with the Call commitments), even if the intermediate pathways are not necessarily via content defined as TVEC. A focus on the user experience will help us to identify intervention and redirection points, as well as better understand the central problem of how algorithmic outcomes may contribute to radicalisation, and ultimately to terrorism and violent extremism.

3. There has been an evolution in the terrorism / violent extremism threat landscape since the Christchurch Call came into effect, including a shift of TVEC to different platforms and the impact of COVID, that needs to be taken into account.

4. We welcome the work that is being progressed in other forums, including the GIFCT Working Groups and GPAI. We note that the GIFCT algorithms and positive interventions Working Group is at the literature review and gap analysis stage and are working towards sharing outputs around the July 2021 Summit.

Better understanding the outputs of recommendation systems based on machine learning

5. The Christchurch Call asked online service providers and Governments jointly to build trusted information-sharing mechanisms to address redirection and algorithmic outcomes.

6. Transparency and explainability around such information-sharing mechanisms is central to their broader legitimacy. There are multiple models. Some come with trade-offs in terms of reactiveness, proprietary data, user data, and potential abuse by Governments or other stakeholders. We don't have all the answers. It's helpful that the Call focuses on the outcomes of algorithms.

7. Transparency may vary between online service providers and information-sharing mechanisms should take this into account. NB: transparency related to other efforts to address terrorist and violent extremist content online is being addressed in a separate workstream.

8. While there are some efforts being made, there is currently a lack of publicly available evaluation data on these recommendation systems, and peer reviewed/multi-stakeholder framework for the evaluation of these initiatives.

9. Further understanding of any tensions within and between regulatory frameworks are needed to ensure that the law supports Christchurch Call efforts.

10. Trusted information-sharing mechanisms for multi-stakeholder engagement are important, and efforts in this area could be improved. The GIFCT does some of this but on a very limited and selective basis.

Positive Interventions, counter-narratives and community empowerment

11. There was significant engagement on this topic, including on identifying ways that the Call Community could work together to develop a broader range of interventions that are both

better evaluated and more effective. Issues around the role of traditional media, and other factors sitting outside of the core user interface were also discussed, and a range of ideas were put forward.

12. Many platforms have programs in place to intervene in user journeys that may lead to terrorist and violent extremist content. These so-called "Redirect" efforts vary in important ways and have achieved mixed levels of success. A thorough, comparative assessment of these programmes will be useful as we expand the aperture for such interventions.

Removal of TVEC and enforcement of policies

13. Many online service providers have robust policies prohibiting TVEC on their platforms. Enforcement of those policies, including through the use of algorithmic methods, is improving but remains imperfect (including with geographic inconsistencies). Some participants noted that perfection may not be possible in this domain. Online service providers should detail the use of automated systems in the enforcement of their rules to the extent possible without disclosing proprietary information or data that would assist nefarious actors in circumventing their systems. Some participants also flagged the importance of explainability of decisions made by algorithms.

14. There was discussion around how online service providers might address the presence of TVEC that inadvertently remains on their platforms, and how that fits with the broader questions around possible amplification.

TVEC - redress mechanisms

15. The practical use of algorithms often requires trading off between accurate detection of TVEC and unintended suppression of content (false positives). Algorithms are imperfect and their practical use often requires trading off broader recall with increased false positives. Given these trade-offs, online service providers should prioritise redress mechanisms for users that have been impacted by content moderation decisions empowered by algorithms and make publicly available more information and data on the action taken with regard to such complaints. Participants also emphasised the importance of human review of decisions taken by algorithms, and appeals processes.

16. As has also been flagged in the Community and Transparency workstreams, there's a need to include more diverse voices in the conversation, including the voices of victims of TVEC, and voices from marginalised communities.

Action Points to be picked up in Work Plan

Consistent with the Christchurch Call commitments, workstream members have identified the following work plan for algorithms and positive interventions to progress the Christchurch Call Commitments.

Actions relate to the three key outcome areas:

1. **Building understanding of recommender algorithms, and user journeys**
2. **Empowering a new generation of community-driven online interventions**
3. **Mechanisms for TVEC removal: Transparency and Redress.**

In developing these actions, we have considered the timeframe (elements are divided into medium (i.e. achievable over the first 6-12 months) and longer term (1-3 years). We have also endeavoured to use the SMART objectives (specific, measurable, attainable, relevant and time-based).

Medium Term Objectives (Achievable in 6 months - 1 year)			
Objective & Action	**Rationale**	**Key stakeholders**	**Evaluation measures**
1. **Building understanding of recommender algorithms, and user journeys** As the Call Community we will devote effort and resources towards better understanding the "user journey" and the role this may play in the broader radicalisation process. We will design a multi-stakeholder process to establish what methods can safely be used, and what information is needed - without compromising trade secrets or the effectiveness of Online Service Providers' practises through unnecessary disclosure - to allow stakeholders to better understand the outcomes of algorithmic processes, and their potential to amplify terrorist and violent extremist content. This could include improving understanding of user journeys online, through looking at questions such as whether terrorists and violent extremists exploit the content recommendation processes of digital platforms to spread or discover terrorist and violent extremist content, their evolving tactics and how this online journey interacts with offline drivers of radicalisation.	Delivering on core Call Commitments Key to addressing underlying questions being asked by Community Members	Call supporting governments and online service providers as well as CCAN and the wider civil society and technical community and academics - working alongside GIFCT and other fora	Re-convene quarterly with workstream participants and other interested members of the Call Community to check in on progress

This will be conducted in an open and transparent way in collaboration with the work of GIFCT and as appropriate in other fora, such as the EU Internet Forum.			
2. Empowering a new generation of community-driven online interventions As the Call Community we will take steps to prevent radicalisation online by assessing the wide range of intervention points on digital platforms where relevant stakeholders can engage to build, empower, and promote healthy communities online and offline and disrupt radicalisation processes using a full spectrum of products, messaging, or community-building tools, while respecting human rights principles (including freedom of expression and privacy). This year: The Call Community, working with the GIFCT, will seek to identify and empower the next generation of digital interventions against radicalisation, working to build a consistent framework for the comparative evaluation of such work. This should look at: the successes and limits of existing efforts to promote counter and alternative narratives online, including for effectiveness, and human rights impact; examining any current gaps identified by the GIFCT CAPPI WG; and an exploration of the full spectrum of available interventions. We will convene a dedicated multi-stakeholder, in-person symposium addressing how user journeys interact with offline drivers of radicalisation and how	Delivering on core Call Commitments Providing tools that can be used in a Call-consistent way	Call supporting governments including government entities and online service providers as well as CCAN and the wider civil society and technical community, and academics - working alongside GIFCT and other fora	As above

these could inform a new generation of community driven online interventions. Governments will work in an open multi-stakeholder context with the community to identify information that could be shared to assist with positive interventions.			
3. Mechanisms for TVEC removal: Transparency and Redress Enforcement by companies of their policies against terrorist and violent extremist content including the identification of terrorist and violent extremist content through machine learning is improving, but will always be limited in understanding the critical elements of context and intent. As per the Call, Platforms are committed to provide efficient complaints and appeals processes and make publicly available information and data on the actions they are taking. This year the Community will host an inclusive discussion on developing a framework to continuously review and improve the efficiency of such measures, and how to support greater transparency and explainability in this area, coordinating with the transparency work stream.	Delivering on Call Commitments Significant demand from Community	Call supporting online service providers, governments, CCAN and wider civil society	As above

Longer Term objectives (Achievable in 2-3 years)			
Objective & Action	Rationale	Key stakeholders	Evaluation measures
1. Building understanding of recommendation systems and user journeys We will continue to develop understanding of the outcomes of algorithms and other processes. Drawing on this understanding, and knowledge from other sources, the Call community will provide guidance and identify best practises to limit the possibility that user journeys facilitate radicalisation. This will help the community to support the development of appropriate and proportionate preventive actions and responses.	As above	Call supporting governments and online service providers as well as CCAN and the wider civil society and technical community and academics - working alongside GIFCT and other fora	Periodic check-ins with the Community
2. Empowering a new generation of community-driven online interventions We will look at progress to help build resilient, healthy and safe communities online. Governments and online service providers will support civil society, and collaboration among all sectors will empower robust and innovative civil society-led initiatives seeking to prevent and push back on terrorism and violent extremism. A consistent evaluation framework for the new generation of positive interventions should be developed jointly by the Call Community working with the GIFCT. Among other things, this should consider effectiveness and human rights impact. The Call Community should also assess Government regulations to understand how they may impact on effective delivery of positive interventions.	As above	Call supporting governments and online service providers as well as CCAN and the wider civil society and technical community and academics - working alongside GIFCT and other fora	

3. Mechanisms for TVEC removal: Transparency and Redress We will look at progressing within a multi-stakeholder forum, guidance and consistency for transparency reports on the key components and criteria for decisions relating to TVEC. This will also be informed by the: • Outcomes of the Call Community discussion on how to support greater transparency and explainability in this area (referred to in the short -term goal above) • the work articulated by the Christchurch Call transparency workstream; • Any related outcomes arising from the new generation of positive interventions and consistent evaluation framework developed collectively by the Call Community and GIFCT (referred to above). The Community should also assess government regulations to understand how they may impact on effective redress transparency reports and the implementation of redress mechanisms by more platforms.		Call supporting governments, online service providers, CCAN and wider civil society. , G	