

# Introducing 2022 GIFCT Working Group Outputs

GIFCT WORKING GROUPS OUTPUT 2022



**GIFCT**  
Global Internet Forum  
to Counter Terrorism

Dr. Erin Saltman  
Director of Programming,  
GIFCT

In July 2020, GIFCT launched a series of Working Groups to bring together experts from across sectors, geographies, and disciplines to offer advice in specific thematic areas and deliver on targeted, substantive projects to enhance and evolve counterterrorism and counter-extremism efforts online. Participation in Working Groups is voluntary and individuals or NGOs leading Working Group projects and outputs receive funding from GIFCT to help further their group's aims. Participants work with GIFCT to prepare strategic work plans, outline objectives, set goals, identify strategies, produce deliverables, and meet timelines. Working Group outputs are made public on the GIFCT website to benefit the widest community. Each year, after GIFCT's Annual Summit in July, groups are refreshed to update themes, focus areas, and participants.

From August 2021 to July 2022, GIFCT Working Groups focused on the following themes:

- Crisis Response & Incident Protocols
- Positive Interventions & Strategic Communications
- Technical Approaches: Tooling, Algorithms & Artificial Intelligence
- Transparency: Best Practices & Implementation
- Legal Frameworks

A total of 178 participants from 35 countries across six continents were picked to participate in this year's Working Groups. Applications to join groups are open to the public and participants are chosen based on ensuring each group is populated with subject matter experts from across different sectors and geographies, with a range of perspectives to address the topic. Working Group participants in 2021–2022 came from civil society (57%), national and international government bodies (26%), and technology companies (17%).

Participant diversity does not mean that everyone always agrees on approaches. In many cases, the aim is not to force group unanimity, but to find value in highlighting differences of opinion and develop empathy and greater understanding about the various ways that each sector identifies problems and looks to build solutions. At the end of the day, everyone involved in addressing violent extremist exploitation of digital platforms is working toward the same goal: countering terrorism while respecting human rights. The projects presented from this year's Working Groups highlight the many perspectives and approaches necessary to understand and effectively address the ever-evolving counterterrorism and violent extremism efforts in the online space. The following summarizes the thirteen outputs produced by the five Working Groups.

### **Crisis Response Working Group (CRWG):**

The GIFCT Working Group on Crisis Response feeds directly into improving and refining GIFCT's own [Incident Response Framework](#), as well as posing broader questions about the role of law enforcement, tech companies, and wider civil society groups during and in the aftermath of a terrorist or violent extremist attack. CRWG produced three outputs. The largest of the three was an immersive virtual series of Crisis Response Tabletop Exercises, hosted by GIFCT's Director of Technology, Tom Thorley. The aim of the Tabletops was to build on previous Europol and Christchurch Call-led Crisis Response events, with a focus on human rights, internal communications, and external strategic communications in and around crisis scenarios. To share lessons learned and areas for

improvement and refinement, a summary of these cross-sector immersive events is included in the 2022 collection of Working Group papers.

The second output from the CRWG is a paper on the Human Rights Lifecycle of a Terrorist Incident, led by Dr. Farzaneh Badii. This paper discusses how best GIFCT and relevant stakeholders can apply human rights indicators and parameters into crisis response work based on the 2021 GIFCT Human Rights Impact Assessment and UN frameworks. To help practitioners integrate a human rights approach, the output highlights which and whose human rights are impacted during a terrorist incident and the ramifications involved.

The final CRWG output is on Crisis Response Protocols: Mapping & Gap Analysis, led by the New Zealand government in coordination with the wider Christchurch Call to Action. The paper maps crisis response protocols of GIFCT and partnered governments and outlines the role of tech companies and civil society within those protocols. Overall, the output identifies and analyzes the gaps and overlaps of protocols, and provides a set of recommendations for moving forward.

### **Positive Interventions & Strategic Communications (PIWG):**

The Positive Interventions and Strategic Communications Working Group developed two outputs to focus on advancing the prevention and counter-extremism activist space. The first is a paper led by Munir Zamir on Active Strategic Communications: Measuring Impact and Audience Engagement. This analysis highlights tactics and methodologies for turning passive content consumption of campaigns into active engagement online. The analysis tracks a variety of methodologies for yielding more impact-focused measurement and evaluation.

The second paper, led by Kesa White, is on Good Practices, Tools, and Safety Measures for Researchers. This paper discusses approaches and safeguarding mechanisms to ensure best practices online for online researchers and activists in the counterterrorism and counter-extremism sector. Recognizing that researchers and practitioners often put themselves or their target audiences at risk, the paper discusses do-no-harm principles and online tools for safety-by-design methodologies within personal, research, and practitioner online habits.

### **Technical Approaches Working Group (TAWG):**

As the dialogue on algorithms and the nexus with violent extremism has increased in recent years, the Technical Approaches Working Group worked to produce a longer report on Methodologies to Evaluate Content Sharing Algorithms & Processes led by GIFCT's Director of Technology Tom Thorley in collaboration with Emma Llanso and Dr. Chris Meserole. While Year 1 of Working Groups produced a paper identifying the types of algorithms that pose major concerns to the CVE and counterterrorism sector, Year 2 output explores research questions at the intersection of algorithms, users and TVEC, the feasibility of various methodologies and the challenges and debates facing research in this area.

To further this technical work into Year 3, TAWG has worked with GIFCT to release a Research Call

for Proposals funded by GIFCT. This Call for Proposals is on Machine Translation. Specifically, it will allow third parties to develop tooling based on the [gap analysis](#) from last year's TAWG Gap Analysis. Specifically, it seeks to develop a multilingual machine learning system addressing violent extremist contexts.

### **Transparency Working Group (TWG):**

The Transparency Working Group produced two outputs to guide and evolve the conversation about transparency in relation to practitioners, governments, and tech companies. The first output, led by Dr. Joe Whittaker, focuses on researcher transparency in analyzing algorithmic systems. The paper on Recommendation Algorithms and Extremist Content: A Review of Empirical Evidence reviews how researchers have attempted to analyze content-sharing algorithms and indicates suggested best practices for researchers in terms of framing, methodologies, and transparency. It also contains recommendations for sustainable and replicable research.

The second output, led by Dr. Courtney Radsch, reports on Transparency Reporting: Good Practices and Lessons from Global Assessment Frameworks. The paper highlights broader framing for the questions around transparency reporting, the needs of various sectors for transparency, and questions around what meaningful transparency looks like.

### **The Legal Frameworks Working Group (LFWG):**

The Legal Frameworks Working Group produced two complementary outputs.

The first LFWG output is about Privacy and Data Protection/Access led by Dia Kayyali. This White Paper reviews the implications and applications of the EU's Digital Services Act (DSA) and the General Data Protection Regulation (GDPR). This includes case studies on Yemen and Ukraine, a data taxonomy, and legal research on the Stored Communications Act.

The second LFWG output focuses on terrorist definitions and compliments GIFCT's wider Definitional Frameworks and Principles work. This output, led by Dr. Katy Vaughan, is on The Interoperability of Terrorism Definitions. This paper focuses on the interoperability, consistency, and coherence of terrorism definitions across a number of countries, international organizations, and tech platforms. Notably, it highlights legal issues around defining terrorism based largely on government lists and how they are applied online.

### **Research on Algorithmic Amplification:**

Finally, due to the increased concern from governments and human rights networks about the potential link between algorithmic amplification and violent extremist radicalization, GIFCT commissioned Dr. Jazz Rowa to sit across three of GIFCT's Working Groups to develop an extensive paper providing an analytical framework through the lens of human security to better understand the relation between algorithms and processes of radicalization. Dr. Rowa participated in the Transparency, Technical Approaches, and Legal Frameworks Working Groups to gain insight into

the real and perceived threat from algorithmic amplification. This research looks at the contextuality of algorithms, the current public policy environment, and human rights as a cross-cutting issue. In reviewing technical and human processes, she also looks at the potential agency played by algorithms, governments, users, and platforms more broadly to better understand causality.

We at GIFCT hope that these fourteen outputs are of utility to the widest range of international stakeholders possible. While we are an organization that was founded by technology companies to aid the wider tech landscape in preventing terrorist and violent extremist exploitation online, we believe it is only through this multistakeholder approach that we can yield meaningful and long-lasting progress against a constantly evolving adversarial threat.

We look forward to the refreshed Working Groups commencing in September 2022 and remain grateful for all the time and energy given to these efforts by our Working Group participants.

**Participant Affiliations in the August 2021 - July 2022 Working Groups:**

| Tech Sector    | Government Sector                                                           | Civil Society / Academia / Practitioners                                                 | Civil Society / Academia / Practitioners                                                    |
|----------------|-----------------------------------------------------------------------------|------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------|
| ActiveFence    | Aqaba Process                                                               | Access Now                                                                               | Lowy Institute                                                                              |
| Amazon         | Association Rwandaise de Défense des Droits de l'Homme                      | Anti-Defamation League (ADL)                                                             | M&C Saatchi World Services Partner                                                          |
| Automatic      | Australian Government - Department of Home Affairs                          | American University                                                                      | Mnemonic                                                                                    |
| Checkstep Ltd. | BMI Germany                                                                 | ARTICLE 19                                                                               | Moonshot                                                                                    |
| Dailymotion    | Canadian Government                                                         | Australian Muslim Advocacy Network (AMAN)                                                | ModusIzad - Centre for applied research on deradicalisation                                 |
| Discord        | Classification Office, New Zealand                                          | Biodiversity Hub International                                                           | New America's Open Technology Institute                                                     |
| Dropbox, Inc.  | Commonwealth Secretariat                                                    | Bonding Beyond Borders                                                                   | Oxford Internet Institute                                                                   |
| ExTrac         | Council of Europe, Committee on Counter-Terrorism                           | Brookings Institution                                                                    | Partnership for Countering Influence Operations, Carnegie Endowment for International Peace |
| Facebook       | Department of Justice - Ireland                                             | Business for Social Responsibility                                                       | Peace Research Institute Frankfurt (PRIF); Germany                                          |
| JustPaste.it   | Department of State - Ireland                                               | Centre for Analysis of the Radical Right (CARR)                                          | PeaceGeeks                                                                                  |
| Mailchimp      | Department of State - USA                                                   | Center for Democracy & Technology                                                        | Point72.com                                                                                 |
| MEGA           | Department of the Prime Minister and Cabinet (DPMC), New Zealand Government | Center for Media, Data and Society                                                       | Polarization and Extremism Research and Innovation Lab (PERIL)                              |
| Microsoft      | DHS Center for Prevention Programs and Partnerships (CP3)                   | Centre for Human Rights                                                                  | Policy Center for the New South (senior fellow)                                             |
| Pex            | European Commission                                                         | Centre for International Governance Innovation                                           | Public Safety Canada & Carleton University                                                  |
| Snap Inc.      | Europol/EU IRU                                                              | Centre for Youth and Criminal Justice (CYCJ) at the University of Strathclyde, Scotland. | Queen's University                                                                          |
| Tik Tok        | Federal Bureau of Investigation (FBI)                                       | Cognitive Security Information Sharing & Analysis Center                                 | Sada Award, Athar NGO, International Youth Foundation                                       |
| Tremau         | HRH Prince Ghazi Bin Muhammad's Office                                      | Cornell University                                                                       | Shout Out UK                                                                                |
| Twitter        | Ministry of Culture, DGMIC - France                                         | CyberPeace Institute                                                                     | Strategic News Global                                                                       |
| You Tube       | Ministry of Foreign Affairs - France                                        | Dare to be Grey                                                                          | S. Rajaratnam School of International Studies, Singapore (RSIS)                             |
|                | Ministry of Home Affairs (MHA) - Indian Government                          | Dept of Computer Science, University of Otago                                            | Swansea University                                                                          |
|                | Ministry of Justice and Security, the Netherlands                           | Digital Medusa                                                                           | Tech Against Terrorism                                                                      |
|                | National Counter Terrorism Authority (NACTA) Pakistan                       | Edinburgh Law School, The University of Edinburgh                                        | The Alan Turing Institute                                                                   |



|  |                                                                                                       |                                                                 |                                                                                                                |
|--|-------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|
|  | Organisation for Economic Co-operation and Development (OECD)                                         | European Center for Not-for-Profit Law (ECNL)                   | The Electronic Frontier Foundation                                                                             |
|  | Office of the Australian eSafety Commissioner (eSafety)                                               | Gillberg Neuropsychiatry Centre, Gothenburg University, Sweden, | The National Consortium for the Study of Terrorism and Responses to Terrorism (START) / University of Maryland |
|  | Organization for Security and Co-operation in Europe (OSCE RFoM)                                      | George Washington University, Program on Extremism              | Unity is Strength                                                                                              |
|  | Pôle d'Expertise de la Régulation Numérique (French Government)                                       | Georgetown University                                           | Université de Bretagne occidentale (France)                                                                    |
|  | North Atlantic Treaty Organization, also called the North Atlantic Alliance (NATO)                    | Georgia State University                                        | University of Auckland                                                                                         |
|  | Secrétaire général du Comité Interministériel de prévention de la délinquance et de la radicalisation | Global Network on Extremism and Technology (GNET)               | University of Groningen                                                                                        |
|  | State Security Service of Georgia                                                                     | Global Disinformation Index                                     | University of Massachusetts Lowell                                                                             |
|  | The Royal Hashemite Court/ Jordanian Government                                                       | Global Network Initiative (GNI)                                 | University of Oxford                                                                                           |
|  | The Office of Communications (Ofcom), UK                                                              | Global Partners Digital                                         | University of Queensland                                                                                       |
|  | UK Home Office                                                                                        | Global Project Against Hate and Extremism                       | University of Salford, Manchester, England,                                                                    |
|  | United Nations Counter-terrorism Committee Executive Directorate (CTED)                               | Groundscout/Resonant Voices Initiative                          | University of South Wales                                                                                      |
|  | UN, Analytical Support and Sanctions Monitoring Team (I267 Monitoring Team)                           | Hedayah                                                         | University of the West of Scotland                                                                             |
|  | United Nations Major Group for Children and Youth (UNMGCY)                                            | Human Cognition                                                 | Violence Prevention Network                                                                                    |
|  | United States Agency for International Development (USAID)                                            | Institute for Strategic Dialogue                                | WeCan Africa Initiative & Inspire Africa For Global Impact                                                     |
|  |                                                                                                       | International Centre for Counter-Terrorism                      | Wikimedia Foundation                                                                                           |
|  |                                                                                                       | Internet Governance Project, Georgia Institute of Technology    | World Jewish Congress                                                                                          |
|  |                                                                                                       | Islamic Women's Council of New Zealand                          | XCyber Group                                                                                                   |
|  |                                                                                                       | JOS Project                                                     | Yale University, Jackson Institute                                                                             |
|  |                                                                                                       | JustPeace Labs                                                  | Zinc Network                                                                                                   |
|  |                                                                                                       | Khalifa Ihler Institute                                         |                                                                                                                |
|  |                                                                                                       | KizBasina (Just-a-Girl)                                         |                                                                                                                |
|  |                                                                                                       | Love Frankie                                                    |                                                                                                                |

# Methodologies to Evaluate Content Sharing Algorithms & Processes

GIFCT Technical Approaches Working  
Group

GIFCT WORKING GROUPS OUTPUT 2022



**GIFCT**  
Global Internet Forum  
to Counter Terrorism

Tom Thorley, GIFCT *In collaboration with:*  
Emma Llansó, Center for Democracy and Technology  
Chris Meserole, Brookings Institution



## Executive Summary

Over the past 12 months, representatives from government, tech, and civil society have come together as part of the GIFCT Technical Approaches Working Group (WG). The group adopted the shared goal of exploring the research questions needed to be addressed to fully understand the intersection of terrorist and violent extremist content, users and content sharing algorithms and assessing the feasibility of a number of methodologies in order to identify the challenges that research of this kind needs to address in order to provide meaningful insights for policy makers.

This report assessed three methodologies focusing on three different research questions and three different disclosure approaches and recommended that GIFCT seek to arrange meetings to address the legal and technical feasibility of specific aspects of two of the methodologies while the third should be rescoped and redesigned to strengthen safeguards to privacy and ensure that the data requested is necessary and proportionate.

The report also concludes that to properly address technical approaches to answer these research questions methodological design must address definitional issues, generalization, privacy & security, a range of human rights and ultimately the impact on terrorism and violent extremism.

Finally, it identifies a taxonomy of research questions that need to be considered to address knowledge gaps in what is known about terrorist and violent extremist content (TVEC) and algorithmic processes.

Working groups are a multistakeholder effort to further discussion on the given topic of the nexus between terrorism and technology. This paper represents a diverse array of expertise and analysis coming from tech, government, and civil society participants. It is not a statement of policy, nor is this paper to be considered the official view of the stakeholders who provided inputs.

## Introduction

How individuals become radicalized to join violent extremist groups or commit violent acts motivated by extremist ideologies has been studied for many years and the role of the internet in this process has been well documented. Many researchers, governments, and nonprofits have “hypothesized the existence of a radicalization pipeline,”<sup>1</sup> linking content-sharing algorithms with radicalization. Whereas recent research has shown that platform efforts to take content quality into account when recommending content has been effective in lowering risk<sup>2</sup>, GIFCT member companies seek to remove terrorist and violent extremist content and have made significant improvements in how they manage content-sharing algorithms – such as YouTube’s 2019 update to its algorithm<sup>3</sup> – since which time recommendations of such material have been “relatively uncommon and heavily concentrated in a small minority of participants who previously expressed high levels of hostile sexism and racial resentment.”<sup>4</sup> However, many questions remain about “borderline content”<sup>5</sup> as well as causality and agency in these complex and dynamic processes. Ultimately, assurance is needed that when adding a feature or technology to the web, the harm it could do to society or groups (especially to vulnerable people) has been considered.<sup>6</sup>

In July 2020, the GIFCT established the Content-Sharing Algorithms, Processes, and Positive Interventions Working Group (CAPPI), made up of representatives from governments, tech companies, and civil society, including academia, practitioners, human rights experts, researchers, and members of the NGO community, who produced a report in July 2021 mapping content-sharing algorithms and processes used by industry.<sup>7</sup>

This paper builds on CAPPI’s initial report to evaluate methodologies for researching this topic, identifies issues that prevent studies using these methodologies from moving forward, and identifies potential pilot studies to be commissioned as a means to further the discussion and improve methodological design. The paper also reflects work done with tech companies to identify the processes and best practices they have in place to guide their engagement in research and ensure responsible and ethical research practices (see Appendix C for details).

The Christchurch Call to Action produced a work plan in May 2021 to “provide impetus and

1 Manoel Horta Ribeiro et al., “Auditing radicalization pathways on YouTube,” Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, <https://doi.org/10.1145/3351095.3372879>.

2 Lewis-Kraus, G. (2022, June 3). How Harmful Is Social Media? The New Yorker. <https://www.newyorker.com/culture/annals-of-inquiry/we-know-less-about-social-media-than-we-think>.

3 “Continuing our work to improve recommendations on YouTube,” YouTube (Blog), January 25, 2019, <https://blog.youtube/news-and-events/continuing-our-work-to-improve/>.

4 Annie Y. Chen et al., “Subscriptions and external links help drive resentful users to,” April 22, 2022, arXiv.Org, <https://arxiv.org/abs/2204.10921>.

5 Amélie Heldt, “Borderline speech: caught in a free speech limbo?,” Internet Policy Review, October 15, 2020, <https://policyreview.info/articles/news/borderline-speech-caught-free-speech-limbo/1510>.

6 “W3C TAG Ethical Web Principles,” World Wide Web Consortium (W3C), May 12, 2022, <https://www.w3.org/TR/2022/DNOTE-ethical-web-principles-20220512/#noharm>.

7 “Content-Sharing Algorithms, Processes, and Positive Interventions Working Group Part 1: Content-Sharing Algorithms & Processes,” GIFCT, July, 2021, <https://gifct.org/wp-content/uploads/2021/07/GIFCT-CAPPI-2021.pdf>.

momentum” supporting call participants to “review the operation of algorithms and other processes that may drive users towards and/or amplify terrorist and violent extremist content.” This work plan focused on “building understanding of recommendation systems and user journeys.”<sup>8</sup>

There is more work to do to understand the agency and causal mechanisms at play that may link radicalization and recommender algorithms and “to fully understand the problems inherent in ‘de-amplifying’ legal, borderline content.”<sup>9</sup> As a result, the potential role of content-sharing algorithms in radicalization and violent extremist recruitment continues to be an issue of focus for GIFCT working groups.

The nature of GIFCT’s working groups is that many perspectives are represented from different sectors, including tech companies, governments, academia, and civil society. In writing this paper we aimed to both seek consensus and highlight the debates and counterpoints to various issues where a consensus position has not been reached.

The methodologies and pilot studies discussed in this paper should not be considered as a commitment to conduct a pilot but rather a commitment to discuss the feasibility of the methodology and how each could be taken forward or redesigned. This is a continuing effort and will be an iterative process.

## Definitions and Descriptions of Key Terms

### Terrorist and Violent Extremist Content (TVEC)

A key piece of the GIFCT Membership criteria<sup>10</sup> is that members must prohibit terrorist and/or violent extremist exploitation of their services and include this explicitly in their publicly-available terms of service or content standards. Just like governments, intergovernmental institutions, civil society organizations, and academics, tech companies often have slightly different definitions of “terrorism,” “terrorist content,” and “violent extremism.” While there is no one globally agreed-upon definition of terrorism or violent extremism, most tech companies in their independent capacity have developed definitions and approaches based on existing resources and in consideration of what will work best based on how their platforms operate.

For example, Meta’s dangerous individuals and organizations policies<sup>11</sup> explicitly prohibits any organizations or individuals that proclaim a violent mission or are engaged in violence to have a presence on their platform. This includes organizations or individuals involved in terrorist activity,

8 Christchurch Call to Action, “Second Anniversary of the Christchurch Call Summit, Joint Statement by Prime Minister Rt Hon Jacinda Ardern and His Excellency President Emmanuel Macron as co-founders of the Christchurch Call,” May, 2021, <https://www.christchurchcall.com/second-anniversary-summit-en.pdf>.

9 Joe Whittaker et al., “Recommender systems and the amplification of extremist content,” Internet Policy Review 10, no. 2 (2021), <https://doi.org/10.14763/2021.2.1565>.

10 “Membership,” GIFCT, January 8, 2022, <https://gifct.org/membership/>.

11 “Meta - Dangerous Individuals and Organizations,” Facebook.com, April, 2022, <https://www.facebook.com/communitystandards/dangerous-individuals-organizations>.

organized hate, mass murder, organized violence, and large-scale criminal activity. This is an approach based on the behaviors of organizations and entities.

Microsoft's standards<sup>12</sup> also prohibit terrorist content. This is defined as material posted by or in support of organizations included on the Consolidated United Nations Security Council Sanctions List that depicts graphic violence, encourages violent action, endorses a terrorist organization or its acts, or encourages people to join such groups. In contrast to Meta's approach, Microsoft takes an approach based on a list defined by an intergovernmental body.

JustPaste.It's Terms of Service<sup>13</sup> prohibits terrorist content, which it defines as content in violation of EU Directives and EU Member State laws on terrorist offenses, or content produced by or attributable to terrorist groups or entities designated by the European Union or by the United Nations. JustPaste.It's definition is a mixture of a legally-based definition (relying on the laws in the jurisdiction where they are based) and a list-based approach similar to the one taken by Microsoft.

How a company defines terrorism and violent extremism relies on a number of different factors, including the legal jurisdictions in which they operate, the function of the relevant platform, and the use cases of their users. They must also seek to build community standards that can be enforced practically by content moderators and so definitions and descriptions must be both practically applicable and easily understood.

Governments, academics, and others also provide definitions of terrorism and violent extremism, and while there are many overlapping aspects of these definitions a consensus has not yet been reached. GIFCT is engaged in ongoing work to provide a definitional framework and supporting material to help aid members in navigating this challenging issue.

However, a company defines terrorism and violent extremism within their community standards, it is clear that GIFCT member companies enforce their own respective policies and conduct their own practices in response to violations of their terms of service or standards such as content removal and account disabling. Once content is identified as Terrorist and Violent Extremist Content (TVEC) under a platform's community guidelines, that content will be removed by the platform.

In addition to the lack of consistency between company definitions, the challenges described above are compounded when considering the legality of what we might consider TVEC. Laws which proscribe against extreme content are diverse, with many countries or international organizations holding different conceptualizations as to what constitutes illegal "extreme," "terrorist," "hateful," etc. content <sup>14</sup>.

.....  
 12 "Microsoft's approach to terrorist content online," Microsoft (blog), June 13, 2017, <https://blogs.microsoft.com/on-the-issues/2016/05/20/microsofts-approach-terrorist-content-online/>.

13 "Terms of Service," JustPaste.It, April, 2022, <https://justpaste.it/terms>.

14 Meserole, C., & Byman, D. (2022, July 19). Terrorist Definitions and Designations Lists: What Technology Companies Need to Know. Royal United Services Institute. <https://rusi.org/explore-our-research/publications/special-resources/terrorist-definitions-and-designations-lists-what-technology-companies-need-to-know/>.

## Borderline Content

Just as there is no standard broadly accepted definition of TVEC, there is no standard or broadly accepted definition of “borderline content.”

However, examining the community guidelines, terms of service, acceptable use policies, and other relevant publications of YouTube, Twitter, Microsoft, and Meta does provide some indications of the kinds of content that can be considered as “borderline.” Both YouTube and Meta refer to borderline content in their transparency materials. YouTube describes borderline content as follows:

**Content that comes close to – but doesn’t quite cross the line of – violating our Community Guidelines.<sup>15</sup>**

While Meta has developed specific categories of borderline content, it too uses similar language in its community standards.<sup>16</sup>

**Types of content that are not prohibited by our Community Standards but that come close to the lines drawn by those policies.<sup>17</sup>**

A working description of borderline content in the context of terrorism and violent extremism could therefore be:

**Content that comes close to violating policies around terrorism and violent extremism and that shares some characteristics of hateful or harmful content.**

However, this description is not practically useful as a definition of borderline content as it does not offer a standard against which to judge an individual piece of content. As the nuances of the policies on each platform and the strategies each can employ to manage this content are different, the description is neither precise nor generalizable. When it comes to measuring the impact of borderline content on radicalization or the impact of algorithms recommending this content to users, this lack of precision and generalizability means that comparative studies across platforms will be highly challenging as will cross-platform recommendations about methodologies for research, development safeguards, or other interventions.

## Content-Sharing Systems

As this paper builds on the work on CAPPI, in it we consider content-sharing systems or “recommender algorithms” in the way defined by their report on Content-Sharing Algorithms & Processes:

.....

<sup>15</sup> “The Four Rs of Responsibility, Part 2: Raising authoritative content and reducing borderline content and harmful misinformation,” YouTube (blog), December 3, 2019, <https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/>.

<sup>16</sup> “Facebook Community Standards,” Facebook, 2022, <https://transparency.fb.com/policies/community-standards/>.

<sup>17</sup> “Content Borderline to the Community Standards,” Meta, September 23, 2021, <https://transparency.fb.com/features/approach-to-ranking/content-distribution-guidelines/content-borderline-to-the-community-standards/>.



**In contrast to search algorithms, recommendation algorithms typically do not share content in response to explicit user input such as a search query, but instead surface relevant and engaging content automatically.<sup>18</sup>**

In their paper “Artificial Intelligence, Content Moderation, and Freedom of Expression,” the Transatlantic Working Group describes recommender systems as “automated tools that present (‘curate’) a selection of content (‘recommendations’) from an abundance of content.”<sup>19</sup>

As well as having a common understanding of what recommender systems are, it is also important to point out that recommendations can be considered a form of content moderation themselves. They are designed to be non-neutral and recommend some kinds of content and “downrank” others in accordance with the company’s terms of service and policies (which in the case of GIFCT members precludes TVEC and often limits borderline content).

## **Vulnerable Users**

As we assess the impact of content-sharing recommendation systems on users, we should pay particular attention to the rights, needs, and challenges of individuals from groups or populations that may be at heightened risk of becoming vulnerable. Vulnerable groups are those that face being marginalized, discriminated against, or exposed to other adverse human rights impacts with greater severity and/or lesser potential for remediation than others.

Vulnerability depends on context, and someone who may be powerful in one context may be vulnerable in another. Examples include:

- Formal Discrimination: Laws or policies that favor one group over another.
- Societal Discrimination: Cultural or social practices that marginalize some and favor others.
- Practical Discrimination: Marginalization due to life circumstances, such as poverty.
- Hidden Groups: People who might need to remain hidden and consequently may not speak up for their rights, such as undocumented migrants.

Examples of vulnerable groups, based on input from BSR as part of the GIFCT Human Rights Impact Assessment, are included in Appendix B (though every case is unique).

## **Research Questions**

In order to evaluate methodologies for researching this topic, it is important that we focus on the particular research questions that need to be answered in this space.

Our working group solicited feedback from participants and members of participating organizations

.....

<sup>18</sup> “Content-Sharing Algorithms,” GIFCT. <https://gifct.org/wp-content/uploads/2021/07/GIFCT-CAPII-2021.pdf>.

<sup>19</sup> Emma Llansó et al., “Artificial Intelligence, Content Moderation, and Freedom of Expression,” Transatlantic Working Group, February, 2020, <https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf>.

to build a long list of questions that could be further explored. We categorized this list, grouping questions based on what is being affected and by what. The full list of 31 research questions considered is available in Appendix A, and the diagram below shows the matrix of questions with an overarching question indicating the kind of effect being explored in each case.

| Affected Party<br>Source of effect     | Content-Sharing Recommendation Systems                                                | Users and User Behavior                                                                                           | Content Engagement and Reach                                                                                        |
|----------------------------------------|---------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------|
| Content-Sharing Recommendation Systems |                                                                                       | What are the characteristics of users that increase the chances that they will be recommended borderline content? | What are the characteristics of borderline content that increases the chances that it will be recommended to users? |
| Users and User Behavior                | What is the impact of Content Recommending System on Users' Behavior?                 |                                                                                                                   | What is the impact of borderline content on users?                                                                  |
| Content Engagement and Reach           | What is the impact of Content Recommending System on the reach of borderline content? | What are the characteristics of users most likely to consume and share borderline content?                        |                                                                                                                     |

To focus discussion and enable robust evaluation of the associated methodologies, we selected three questions representing different effects, research methodologies, and approaches to disclosure of information:

- What users are most likely to have borderline content recommended to them?
- What are the effects of recommender systems on platform users' attitudes towards TVEC?
- How is TVEC and borderline content that is ultimately moderated recommended by content-sharing recommender systems before and after moderation takes place?

We also consider that perhaps key to this set of issues is the question, "What is the impact of borderline content on users?" There is still significant debate in this area and further research is required to show the causal links and factors that affect any impact that both TVEC and borderline content have on users.

## Question 1

*What users are most likely to have borderline content recommended to them?*

### Context

A common concern about recommendation algorithms is that they expose users to borderline extremist content they would not otherwise consume. As one analyst put it, recommendation algorithms seem "to have concluded that people are drawn to content that is more extreme than what they started with – or

to incendiary content in general.”<sup>20</sup>

Yet the extent to which this is a true description of recommendation algorithms remains unknown. Although some research has been carried out on the relationship between recommendation algorithms and extremism, most of that research has focused on the impact of recommendation algorithms on user behaviors and attitudes, rather than the impact of user attributes on the behavior of recommendation algorithms. A notable exception is seen in Annie Chen’s work,<sup>21</sup> which is limited to just participants in the U.S. using a subset of browsers and devices to access YouTube. As a result, we have a very poor understanding of what kinds of users are most likely to be recommended borderline content in the first place.

## Methodology and Data Requirements

That there has been little research on this question is not surprising. Most platforms do not require users to provide demographic information at sign up. As a result, for researchers to develop a clear understanding of which users are most likely to be recommended borderline content in a context without any access to platforms’ internal data, they would need:

1. The ability to manipulate the behavior of synthetic user accounts
2. The ability to observe the content recommended to those accounts

Suffice to say, it is neither feasible nor desirable for researchers to employ either capability with respect to the accounts of real users. Although the ability to vary user behavior and attributes makes it possible to draw strong inferences about any corresponding variation in recommended content, researchers should not exercise control over the behavior and demographic information associated with real user accounts.

An alternative approach to consider would be to look at a content-centric approach rather than user-centric. This would require access to a limited dataset of identified borderline contents and the conditions in which they were recommended to a given typology of users. Such an approach would preserve the privacy and anonymity of the users involved in the dataset (cohorts) without forcing the platforms to provide sensitive information while revealing meaningful insights into real users and content journeys. While the goal of preserving user privacy and anonymity when data is shared has merit, the generation of such datasets requires substantial processing that could implicate applicable legal requirements. For example, the EU General Data Protection Regulation (“GDPR”) imposes requirements on the use of personal data by both platforms and researchers. While researchers’ obligations may not be as heavily implicated by a content-based approach, the compilation of such data requires companies to process personal data and thus to comply with various requirements, including those related to transparency, purpose limitation, and using an appropriate lawful basis for processing. Such compliance requires a case-by-case analysis of the impact of the proposed

.....  
<sup>20</sup> Zeynep Tufekci, “YouTube, the Great Radicalizer,” The New York Times, March 10, 2018, <https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html>.

<sup>21</sup> Chen et al., “Subscriptions and external links.”

research on affected individuals, as well as a determination of the appropriate safeguards required. The failure to comply or impose such safeguards can have significant legal and reputational consequences for platforms. Without first working to assess what processing of personal data is involved and to understand the associated legal requirements and risks, platforms may not be willing to undertake such a pilot study altogether.

While this approach may shed some light on the correlation between cohorts of users and recommendations made, given the legal and ethical challenges, a causal relationship would be challenging to prove.

To address the question, “What users are most likely to have borderline content recommended to them?,” we evaluated a methodology that seeks to avoid both manipulating real users and the legal and ethical challenges faced by a content-centric approach. This involves the use of simulated accounts created and controlled by researchers. Much of what little we do understand about recommendation algorithms relies on this approach: the Wall Street Journal, for instance, created and controlled 100 synthetic accounts on TikTok<sup>22</sup> to better understand the different types of content different types of users were exposed to.

Researcher-controlled accounts could help understand what types of users are most likely to be recommended borderline content. Researchers could generate accounts whose behaviors and demographic information (where applicable) are designed to match specific groups of interest, and then could observe and record what types of content are recommended to them. For example, researchers could create an account on a social media network that initially follows a prominent liberal or conservative media personality, and then record whether that account is more likely to be exposed to borderline content from far-left or far-right movements. By comparing the content recommendations the algorithm makes for different kinds of users, researchers can start to piece together which types of users are most likely to be recommended borderline extremist content.

It is also important to note that many platforms do not require demographic information at sign up. In these cases, demographic data that is assigned to accounts is normally inferred from user behavior to enable targeting of advertising. This means that understanding which users are most likely to be recommended borderline content is only possible with reference to user (real or synthetic) behavior. Segmenting data by user demographics is in many cases actually segmenting by patterns of behavior.

## Data Disclosure

Researcher-generated accounts can be created and controlled either manually or programmatically. In the former case, researchers would create and manipulate user accounts on their own, engaging and interacting with a given platform just as a human would. The main accommodations this would require on the part of the platform are the establishment of a notification mechanism (so that researchers can specify to the platform which accounts are fake) and potentially an exception to the

.....

<sup>22</sup> “Inside TikTok’s Algorithm: A WSJ Video Investigation,” Wall Street Journal, July 21, 2021, <https://www.wsj.com/articles/tiktok-algorithm-vid-ee-investigation-11626877477>.

platform's terms of service (for platforms that ban inauthentic behavior).

Researcher-generated accounts that were automated or programmatically controlled would also require access to the platform's underlying API, potentially including endpoints that are not otherwise publicly available. Permissioning researchers to automate account behaviors would enable researchers to more efficiently explore the potential state space of a given environment and develop a richer understanding of which behaviors and profiles are most likely to lead to the recommendation of borderline content.

To further isolate this researcher from users, researchers could generate and control user accounts within a simulation of the platform rather than the platform itself. As long as the distribution of user profiles, behaviors, and contents on the simulated platform is identical to that of the actual platform, interacting with the simulated environment would help design an evaluation procedure in a controlled environment before actually playing it on the real platform. Although the costs incurred with developing a simulated environment are non-trivial, some major platforms have already developed them for exactly this form of exploratory research.<sup>23</sup>

By creating and controlling simulated user accounts, researchers can develop a first understanding of which user behaviors and attributes are most likely to lead a recommendation algorithm to expose users to borderline extremist content. Further, by relying only on manual manipulation, "black-box" API access, or targeted internal data sharing (i.e., the actual typology of cohorts of users who were recommended a set of identified borderline content), they can develop that understanding without compromising the privacy of real users or disclosing proprietary information about a given algorithm's underlying architecture and performance.

## Initial Ethics Risk Assessment

Based on the initial assessment against the research framework used for GNET research, this methodology's ethical risk is assessed as minimal. However, a comprehensive evaluation of the potential ethical considerations such as these and the limitations on the effectiveness of potential mitigations at the outset is critical to respecting individuals' rights to informational self-determination.

## Limitations and Design Considerations

### Generalizability from Synthetic to Real Users

Relying on researcher-generated accounts raises questions about the generalizability of any findings to real-world use cases; researchers would need to provide an argument for why the user behaviors and attributes they simulate correspond to those of real users of interest. Researchers may need to be given special permissions to create and manipulate inauthentic accounts and potentially provided with special access to the platform's underlying API or synthetic environments. This challenge applies doubly when it comes to evaluating research with synthetic accounts in synthetic environments. To

.....  
 23 "WES: Agent-based User Interaction Simulation on Real Infrastructure," Meta Research, April 29, 2020, <https://research.facebook.com/publications/wes-agent-based-user-interaction-simulation-on-real-infrastructure/>.

be actionable, changes proposed in the findings of this research would first need to correlate with improvements in the real environments. While this does not preclude this kind of research, it does suggest that relying exclusively on synthetic environments is insufficient to answer the research question here.

## Synthetic Environments

A key requirement is the availability of synthetic environments that can be provided with sufficient access control to external researchers. To date, although various synthetic environments have been developed by platforms, no environments that could be used for this specific research have yet been identified. Identifying suitable synthetic environments is the key next step in moving this work forwards.

Moreover, the methodology requires the existence of technical and operational functionality that may exist on some platforms but not others. For example, using this methodology to study a particular platform's recommendation algorithm would effectively require that the platform already had an API in place that could enable such research. Many platforms do not yet provide such research APIs, so this methodology would require the platform to develop and make available such a functionality first.

## Pace of Change

For large tech companies, the pace of change poses challenges to creating a representative synthetic environment. Given the volume of changes seen to a platform's code base on a regular basis, one solution is to use "diff batching" – basically grouping code modifications together in intelligent ways to identify the cluster of related changes that contribute to an effect observed. More work is needed on smarter clustering techniques that group code and infrastructure modifications in order to understand whether and how changes have occurred in the live environment that may affect the representativeness of a synthetic test environment. For experiments using researcher-generated accounts in a live environment, it would be vital to track any changes that occur to the platform's code base during the course of the study, as such changes could materially impact the results.

## Scale and Complexity of Synthetic Accounts

The scale of the research required to show causal relationships is potentially quite significant. Although synthetic environments should be able to perform at similar scales to the platforms themselves, creation and manipulation of synthetic accounts on these systems need to be designed in such a way that it is manageable for researchers.

In addition, the synthetic accounts would need to be generated in such a way that they can represent the social graph of the users on the live platform. This synthetic graph is a complex research question in its own right and needs significant investment.

The synthetic accounts must also be able to simulate realistic social interactions and avoid unnatural



behavior. To speed up research, synthetic accounts could be designed to interact faster than any human ever could; however, it is unknown what the impact of such inauthentic behavior may be. Tackling this may also be constrained by the need for user privacy, as realistic behavior needs to be learned or at least measured against something.

## **Synthetic Accounts Interacting With “Real” Users**

Where synthetic environments are not available, the use of automated or researcher-controlled accounts to engage and interact would not exist in isolation. Instead, they would likely involve interactions with human subjects who may not be aware that they are interacting with a fake account. This is a much more ethically challenging scenario than synthetic accounts interacting with each other in a synthetic environment and would require significant and robust scrutiny and assessment, particularly given the focus of this research on TVEC.

While the methodology discusses establishing a notification mechanism to identify fake accounts, such notification may not be sufficient to adequately inform individuals that the account is being used for research purposes.

Moreover, many platforms prohibit the utilization of fake accounts altogether in their terms of service. For those who do not have terms of service that prohibit the utilization of fake accounts, users may, for example, think such accounts are only used to evaluate technical aspects of the platform rather than to observe their threads, posts, and comments and the extent to which such material is recommended.

## **Adversary Activity Using Synthetic Accounts**

Most platforms restrict the creation and manipulation of synthetic accounts for two reasons – it can allow for both greater coordination of inauthentic behavior on the platform and it can offer greater opportunity for adversarial learning by malicious actors. Both may degrade the experience of authentic users on a given platform. Any API for automating synthetic accounts would thus need to be both permissioned (so that only legitimate researchers have access to it) and monitored (so any attempt to reverse engineer or game the platform’s recommendation algorithms for malicious purposes would be flagged).

## **Next Steps**

Although similar synthetic environments have been built and made available for research in the past, these solutions are expensive and time consuming. Given the limited generalizability of this approach and the limited insights they could provide, it might not justify the expense in the context of such a trial. Ultimately this is an engineering challenge rather than a research challenge, and tech companies should engage in meaningful discussions to identify efficient approaches to a pilot that could provide insightful results on real-world behaviors in a technically viable way. This could lead to developing synthetic environments, which may require significant technical investment, or exploring the policy and ethical challenges of using synthetic accounts on the live platform.

GIFCT should seek to identify a research team with the capacity to further the design and implementation of this project no later than October 2022.

GIFCT should arrange meetings between specific GIFCT member companies (including relevant technical experts) and the research team to explore the technical viability of this project, with a view to reaching a decision no later than the end of 2022.

## Question 2

*What are the effects of recommender systems on platform users' attitudes towards TVEC?*

### Context

There are many existing studies of the effects of social media recommender systems on users. The great majority of these have been conducted “externally” (i.e., not by platforms,) using public APIs provided by companies, data from browser loggers installed by volunteers, or other methods (e.g., simulations of social media users or full social media systems or larger-scale population studies). A review of the available methods is given in Knott et al.<sup>24</sup> However, the results of existing studies are not clear-cut because external methods of studying recommender system effects all have empirical shortcomings, such as confounding variables, sampling problems, and API limitations. A fundamental problem is that external methods do not allow the testing of causal hypotheses about recommender system effects because they cannot intervene in the recommender system placed before users. Knott et al. argue that by far the best methods for studying effects of recommender systems are those used by companies themselves to develop and evaluate their own systems. The proposed pilot project involves working with one or more social media companies to extend their own methods for studying the effects of their recommender systems on users. The pilot study would examine the effects of the recommender system on users' relationship toward TVEC content. This pilot study was originally recommended by the Global Partnership on AI (GPAI)'s project on Social Media Governance in November last year at the GPAI summit.

### Methodology and Data Requirements

Companies currently study recommender system effects in randomized controlled trials that present different recommender system experiences to different user groups and then look for differences in the behavior of users from different groups (see e.g., Shani and Gunawardana as well as Brost et al.)<sup>25</sup> There is some variance in these methods. Some trials are conducted “online” in the form of classic A-B tests that compare different versions of the recommender system. In some companies,

24 Alistair Knott et al., “Responsible AI for Social Media Governance: A proposed collaborative method for studying the effects of social media recommender systems on users,” Global Partnership on Artificial Intelligence, November, 2021, <https://gpai.ai/projects/responsible-ai/social-media-governance/responsible-ai-for-social-media-governance.pdf>.

25 Guy Shani and Asela Gunawardana, “Evaluating recommendation systems,” in Recommender Systems Handbook, eds. F. Ricci et al., (Cham: Springer, 2011), 257–297; Brian Brost et al., “An improved multileaving algorithm for online ranker evaluation,” in Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (July, 2016): 745–748.

trials are also conducted offline, for instance in so-called “multi armed bandit” simulations (see Bottou et al.).<sup>26</sup> Companies also differ in the granularity of their studies. Some companies compare the effects of small changes to the recommender system; others make broader-grained comparisons between “recommender system” and “no recommender-system” groups (as in the study by Huszár et al. of Twitter users).<sup>27</sup>

All these methods could be adapted or extended to study how recommender systems affect users’ attitudes towards borderline content and TVEC. The adaptations involve deploying additional metrics measuring aspects of user behavior that can be used as proxies to measure attitudes towards borderline/TVEC. For example, for each user in the study, the number of searches for borderline/TVEC the user makes could be counted, or the number of times the user engages with TVEC or borderline content (as identified using the companies’ own methods). The user’s engagement with other categories of content seen as contributing to the development of extremism could also be measured (such as hate speech or misinformation). Many companies impose bans on content of this kind and deploy methods for identifying it; the pilot study could readily use these in-house methods, so it uses definitions companies are already working with. But the study could also use publicly-defined metrics used in external studies of recommender systems, such as Brady et al.’s definition of “moral-emotional words”<sup>28</sup> or Rajthe et al.’s definition of “out-group language.”<sup>29</sup>

As different companies will have different in-house methods for measuring the most relevant aspects of user behavior, we suggest designing individual pilot studies with different companies, using the most appropriate metrics available within each company. The choice of metrics is a matter for discussion with companies.

Here are two possible forms for pilot studies to develop:

- A pilot project with social media could use the methodology of the recent study of Twitter users by Huszár et al. which took advantage of testing being done around algorithmic, adapted to focus on users’ exposure to TVEC-related material.<sup>30</sup> The null hypothesis tested here is that there is no difference between “recommender system” users and “no recommender system” users as regards their experience of TVEC (as measured by chosen metrics).
- A pilot project with a social media company could draw on companies’ online A-B testing paradigms or companies’ offline “bandit-style” studies, again adapted to focus on users’ exposure to TVEC-related material. The null hypothesis tested here is that users’ experience of TVEC (as measured by chosen metrics) does not depend on the version of the recommender

26 Léon Bottou et al., “Counterfactual reasoning and learning systems: The example of computational advertising,” *Journal of Machine Learning Research* 14, no. 11 (2013): 3207–3260.

27 Feenc Huszár et al., “Algorithmic amplification of politics on Twitter,” *PNAS* 19(1):e2025334119, December 21, 2021, <https://www.pnas.org/content/119/1/e2025334119>.

28 William J. Brady et al., “Emotion shapes the diffusion of moralized content in social networks,” *Proceedings of the National Academy of Sciences* 114, no. 28 (2017): 7313–7318.

29 Steve Rathje, Jay J. Van Bavel, and Sander van der Linden, “Out-group animosity drives engagement on social media,” *Proceedings of the National Academy of Sciences* 118, no. 26 (June 23, 2021): e2024292118, <https://doi.org/10.1073/pnas.2024292118>.

30 Huszár, “Algorithmic amplification.”

system they are given.

In each case, the metric for “success” in the study is simply that the study uses a methodology appropriate for testing the null hypothesis. Whether the null hypothesis is supported or rejected is not relevant to success as the goal is simply answering the question.

Given our current understanding of the scale and timeline of any effects we are seeking to identify, it may be necessary to collect these metrics over a significant period of time in order to produce meaningful research results.

## Data Disclosure

We recommend the results of each pilot study be disclosed in the form of a scientific report that describes three things:

1. The form of the randomized controlled trial/study: that is, how it divides users into groups whose feeds are curated by different methods (e.g., by different recommender system versions);
2. The metrics deployed to measure the behavior of users, including (i) definitions of the relevant categories of content, (ii) how these align with the company’s own definitions and methods, (iii) full distribution of the attitudes of users towards borderline/TVEC in each group of users, (iv) how for the users who ultimately engage with TVEC their journey was influenced by a recommender system; and (v) how the company’s content moderation system impacts on the study’s results; and
3. The results: that is, the differences (or lack thereof) between user groups in relation to the metrics measured.

This approach can be used to both provide sufficient transparency and assurance on the effects of algorithms while also ensuring it does not compromise proprietary information, intellectual property, or user privacy.

Since the pilot studies report high-level aspects of user behavior and report aggregate measures within large user groups, we do not anticipate any possible issues relating to user privacy, though given the fact that this research is focused on violent extremism, access to user data could disclose illegal or harmful activity.

## Initial Ethics Risk Assessment

Based on the research framework used for GNET research, based on our initial assessment this methodology’s ethical risk appears high. As framed, participants will take part in the study without additional consent being sought and which could reveal illegal or harmful activity due to the nature of the research.

However, exactly what the nature of consent required is and what counts as disclosure of data need

to be further explored. The intent of this methodology is not to disclose any user data beyond the boundaries of a company, but to bring trusted researchers into the company to conduct research within the company. Whether allowing access to these researchers counts as a disclosure will depend on the specific language in any research agreement, term of service of the platform in question, and the status of these researchers as contingent workers in a given company. Contingent on the outcomes of further work to refine these issues, the risk assessment may fit the criteria for a low-risk methodology.

## Limitations and Design Considerations

### Offline Versus Live Data

Given that companies remove TVEC as soon as they find it, the study may be limited to “offline” datasets, in which TVEC is identified retrospectively, provided these offline datasets contain enough information to reconstruct user journeys. Understanding whether this influences the data or effects identified will be critical to understanding and assessing if this methodology can actually advance the understanding of how recommendations shape users’ attitudes.

### Data Sparsity

Only a small proportion of users seek or engage with actual TVEC. If the study measures engagement with material that is “on the pathway towards” TVEC (for instance, borderline content), data sparsity is less of a problem, because there is more of this material. But what these measures tell us about users’ engagement with actual TVEC is more open to debate. These issues are discussed in more detail in Knott et al., who argue they are resolvable.<sup>31</sup>

### Reliance on in-house methods

Where companies study the effects of different recommender system experiences, the existence of such testing does not mean that these testing methods could be adapted or extended to test user attitudes for any purpose. The mechanisms and granularity of this testing varies and are often designed particularly in the context of product improvement. The extent to which such systems could be adapted to study users’ attitudes towards TVEC, if at all, should be critically examined rather than merely assumed.

Companies’ internal processes for studying the effects of different recommender system experiences are also subject to internal controls that should be acknowledged before preparing a pilot study methodology. For instance, companies may adopt a code of conduct or guidelines with detailed information on topics such as collecting consent and how to approach researching certain topics or user types. They may similarly use standardized consent forms and information sheets that would allow them to follow a set template for each study with consistent language. Such controls are implemented to address ethical and legal concerns across projects. If the pilot study conflicts with a

31

Knott et al., “Responsible AI for Social Media Governance,” see Sections 5.4–5.7.

company's own compliance protocol, this will significantly impact the ability to conduct and release research data without violating existing company protections (or even applicable law). Further illustrative details of these processes are outlined in Appendix C.

A first step in this process would be to directly discuss with companies the technical feasibility of such adaptations to existing processes and engage in an informed debate about whether existing methods can be adapted and what internal controls are in place.

## Legal Considerations

Following the study of Kramer et al., questions of where to draw the line in terms of ethical approaches to A/B testing were raised.<sup>32</sup> In particular, the "importance of informed consent in Internet research ethics" was seen as critical to "protect the basic human rights" of users that are (wittingly or unwittingly) involved in studies.<sup>33</sup> We appreciate the need for a special consenting process for a study like that of Kramer et al., which introduces a new experimental manipulation of recommender systems that changes the experience of users. However, the methodology in this proposed pilot expressly makes use of the manipulations companies already make to recommender systems for the purposes of developing and testing their algorithms: it will not further alter the experience of users in any way. Companies' own experiments with recommender systems are clearly permitted by their existing terms of service. This proposed pilot therefore sits in an interesting new position in relation to consent and raises questions that require further discussion.

Particularly given the restrictions on processing personal data under the GDPR and other applicable privacy laws, studies using the outlined methodology may merit obtaining written consent from every study participant, informing them of how their data will be used and shared and authorizing the individual's participation with full awareness of the purpose, benefits, and risks involved. Naturally, platforms will need to ensure that they would not be releasing more data than was necessary for the purpose of the pilot in accordance with their obligations to minimize the amount of personal data they process. Gathering consent from every user of a platform for an individual research project would be unfeasible and so further explorations with platforms to identify approaches to manage these challenges.

## Next Steps

Further refinement of this methodology is required to adequately address the approach to consent and consider the viability of the solution given the data sparsity. The fact that internal studies may have been conducted using a given methodology does not mean that a more responsible and ethical approach to this research should not be sought. Once these issues are resolved, a pilot study following a similar methodology to the above could be viable on a limited basis.

.....  
 32 Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks, PNAS, February 6, 2014, <https://www.pnas.org/doi/abs/10.1073/pnas.1320040111>.

33 Catherine Flick, "Informed consent and the Facebook emotional manipulation study," Research Ethics 12, no. 1 (August 11, 2015): 14–28, <https://doi.org/10.1177/1747016115599568>.

While some companies implement specific data policies for teams carrying out user research to address relevant data protection law and compliance, not all do or will. In tandem with developing a research methodology, researchers should work with companies to understand the full scope of legal obligations implicated. This includes assessing how the personal data that is produced during or from user-research activities interact with these obligations as well as how the obligations may in turn impact how such data is stored and handled.

GIFCT should seek to arrange meetings between specific GIFCT member companies and the research team to discuss technical aspects of this project (for instance, appropriate metrics for measuring users' attitudes towards TVEC), with a view to reaching a decision no later than the end of 2022.

GIFCT should seek to arrange meetings between specific GIFCT member companies and the research team to discuss legal aspects of this project (relating to privacy and consent) with a view to reaching a decision no later than the end of 2022.

## Question 3

*How is TVEC and borderline content that is ultimately moderated recommended by content-sharing recommender systems before and after moderation takes place?*

### Context

When tech companies take action on either TVEC or borderline content, because it was deemed to be subject to their content policies, it may be the case that this content had previously been recommended to a user before being flagged for moderation actions – or continued to be recommended after such actions occur (in the case that content was not fully removed from the platform). It is assumed that moderation is effectively limiting the degree to which this content is being shared on platforms, but it is unknown if and how moderation actions feed back into recommender systems.

The details of how content is recommended are typically internally focused, so there is little existing research directly addressing this question. External research by Gerrard investigated users' behavior to circumvent content moderation when the signals that are being moderated (e.g., specific hashtags) are distinct from the harmful content itself, allowing user-obfuscated content to continue to be both shared and recommended by algorithms.<sup>34</sup> However, this external research is unable to characterize the broader systemic relationship between moderation and recommendations of the same harmful content.

### Methodology and Data Requirements

Transparency reports by companies provide valuable insight into how TVEC and borderline content

.....  
 34 Ysabel Gerrard, "Beyond the hashtag: Circumventing content moderation on social media," *New Media & Society* 20, no. 12 (May 28, 2018): 4492–4511, <https://doi.org/10.1177%2F1461444818776611>.



is moderated. While information is given about content moderation actions for various harm types, how that content moderation intersects with recommendations remains unclear. Trust is also a significant factor in this area and transparency reporting relies on companies having the trust of wider stakeholder groups such as governments and civil society. This methodology suggests expanded transparency reports and accompanying data segmented by the two types of content that is applicable in these cases: TVEC and borderline content (as described above).

This segmented analysis presupposes data that can be used to answer the research question. The pilot study could involve assessing the extent to which TVEC and borderline content is engaged with up until the point of moderation. This would involve analyzing the distribution of timestamps of recommendations and user engagements with the content until it is moderated, as well as what the distribution of viewership, time delay, and reach of the moderated content is for violating content before it is moderated. Such analysis may already be technologically feasible on certain platforms; for example, YouTube reported in 2019 that “over the last 18 months we’ve reduced views on videos that are later removed for violating our policies by 80%.”<sup>35</sup> This pilot suggests providing further granularity over a regular reporting period by looking specifically at content moderated for violating TVEC and borderline content policies. Additional analysis could include identifying the forms of moderation, as some companies may “downrank” rather than remove content in the scope of the pilot. To assess the extent to which TVEC and borderline content is engaged with after downranking, analysis would involve measuring the effects of downranking over seven, 30, and 90 days.

Moreover, for each moderated TVEC and borderline content, analysis could include how the content was moderated, why it was moderated, how much time passed between the initial post and moderation, and whether moderation was automated or resulted in flagging for human review. Analysis could also be undertaken to measure user engagement with the content by assessing how much it was accessed directly through a shared or sent link, how many users engaged with the content on their feeds (including time-based timelines or news feeds, automatically curated feeds like Instagram’s explore page or Twitter’s trending terms page), and how many impressions the content received from search results.

## Data Disclosure

The data requirements outlined above should be disclosed both as periodic reports of aggregated/summary data and as anonymized raw metadata. The latter will facilitate the compilation of third-party aggregated data (e.g., by CSOs/think tanks/governments) which can verify tech companies’ summaries.

Anonymized raw data would be made available to trusted third parties (CSOs/think tanks/governments) as part of regular, independent audits of the recommendation system to verify aggregated data and conduct tests of the algorithms.

The periodic reports and anonymized raw metadata could be disclosed (a) in a dedicated forum

.....

35 YouTubeInsider, September 3, 2019, <https://twitter.com/YouTubeInsider/status/1168876004716138496?s=20&t=gxxdN88P6StAHB3jJZROEw>.

created by the GIFCT for access by CSOs, governments, and affiliated researchers meeting appropriate human rights criteria and having undergone appropriate ethics reviews for each study, (b) directly to state agencies (e.g., regulators) and named third-party researchers and CSOs, and (c) publicly where possible, with appropriate anonymization and other protocols in place to ensure compliance with privacy regulations. This would not only facilitate wider scrutiny of the reports/data by third-party researchers but will also provide a greater impetus for tech companies to address any evidence of TVEC content and borderline content being algorithmically recommended on their platforms.

All data should be limited to anonymized metadata to mitigate potential breaches of privacy and misappropriation by rogue states.

## Initial Ethics Risk Assessment

Based on the initial assessment against the research framework used for GNET research, this methodologies' ethical risk is assessed as high.

Data is used from participants who will take part in the study without their explicit consent and could disclose illegal or harmful activity due to the nature of the research. Unlike where a user's content may be shared with a researcher-controlled account, as in the methodology outlined as part of Question 1, the collection and release of raw data associated with TVEC systematically increases such unexpected collection and observation of user content. Even where such content may be publicly available, users may have intended them for a limited audience or within a specific context. Focusing on metadata mitigates these challenges to some extent, but further work is required to define the specific data that would be needed to address the research question and whether this contains content and/or personally identifiable information.

The ongoing sharing of data with a third party could also have long-term chilling effects on information sharing, particularly where users have not been asked for their consent. This risk increases further if the scope of sharing includes non-public materials that may be assessed for TVEC, such as users' search histories or other indicators of the types of content a user consumes.

## Limitations and Design Considerations

### Infrastructure and Data Required

This methodology calls for extended transparency reporting and disclosure of raw data that answers specific questions. However different companies have different systems for managing data and may not have sufficient technical infrastructure to address these questions.

For example, while the pilot study methodology acknowledges that content may be "downranked," companies may not specifically identify such content separately. Instead, their recommender systems may simply engage in a continual process of promoting relevant content. As a result, tasks such as identifying the average time delay of content that was "downranked" before it received

such treatment may not be feasible and may not reflect existing data collection and retention by companies.

Similarly, companies may not maintain engagement and moderation metrics across different demographic information about the poster-consumer relationship in the context of TVEC. While some of this data may be more reasonably generated, such as approximating geographical location, other data may require companies to attempt to collect or infer information regarding users' characteristics that they otherwise would not maintain. In some instances, companies may not have sufficient data to generate such inferences.

### Data Aggregation and Anonymization

While the description of this methodology notes that transparency reports can provide insight into content moderation practices, the sharing of raw data either as part of regular independent audits or public releases may raise conflicts with privacy protections under the law. A key consideration will be what sort of raw data is released, and in what state of anonymization. Content data itself is likely to include personal data in ways that are difficult or impossible to anonymize (such as an individual posting a picture of their own face, or sharing their home address in the audio of a video clip). While the use of metadata can help to reduce the degree of personal data involved, such metadata may still be considered personal data under privacy regulatory frameworks depending on the level of identifiability. The use of anonymization techniques may similarly be effective, though the extent and duration of this is unclear.

Data released as part of transparency reporting by platforms is typically metrics relating to moderation activities taken against content that violates policies.<sup>36</sup> This level of aggregation, while useful for transparency around platforms moderation practices, is not sufficient to answer the research question, and so further work is required to understand the level of aggregation and anonymization that would be necessary to protect privacy while still being sufficient to address the research question.

Anonymized and/or aggregated data means gathering information relating to the users of a platform in such a way that "the data cannot identify" the users.<sup>37</sup> Techniques such as differential privacy and Secret Sharing for Private Threshold Aggregation Reporting are practical, privacy preserving approaches that have been shown to be reliable.<sup>38</sup>

However, as datasets grow larger, the extent to which anonymization and aggregation can be effective may shift and comprehensive risk assessment is required to ensure that users cannot in fact be identified and information cannot be joined with or have context added to reverse the effects of

.....  
36 For examples of these reports, see transparency section of GIFCT's resource guide: <https://gifct.org/resource-guide/#row-trans>.

37 "Anonymized and/or Aggregated Data Definition," Law Insider, 2022, <https://www.lawinsider.com/dictionary/anonymized-and-or-aggregated-data>.

38 "GitHub - google/differential-privacy: Google's differential privacy libraries," GitHub, May 17, 2022, <https://github.com/google/differential-privacy>; Alex Davidson et al., "STAR: Secret Sharing for Private Threshold Aggregation Reporting," arXiv.Org, September 21, 2021, <https://arxiv.org/abs/2109.10074>.

anonymization. To compound these risks, there is “difficulty in determining anonymity, as it depends on criteria that could change according to technical advances or even by the specific analysis conditions.”<sup>39</sup>

To mitigate these risks, other mechanisms must be combined with anonymization and aggregation to improve privacy protection.<sup>40</sup> While the methodology incorporates references to safeguards in the context of such sharing or release, they cannot be considered in the abstract or passingly acknowledged. As noted above, such safeguards are often a critical element in privacy laws. Prior to the development of a pilot study, researchers must work with companies to assess what specific mitigations may be required in light of a particular research question and design their study to incorporate those safeguards from the outset.

Furthermore, there is significant anti-Islamic bias in the counterterrorism field.<sup>41</sup> According to the proceedings of the 1st Conference on Fairness, Accountability, and Transparency,

**Privacy literature seldom considers whether a proposed privacy scheme protects all persons uniformly, irrespective of membership in protected classes or particular risk in the face of privacy failure. Just as algorithmic decision-making systems may have discriminatory outcomes even without explicit or deliberate discrimination, so also privacy regimes may disproportionately fail to protect vulnerable members of their target population, resulting in disparate impact with respect to the effectiveness of privacy protections.**<sup>42</sup>

As a result, every effort must be made to ensure that biases are both understood and mitigated.

### Adverse Incentives

In each of the cases outlined above – and particularly if the proposed data disclosures are projected to significantly influence future regulation – there is a risk that tech companies will tailor their recommendation systems to produce the best possible results for the reported metrics at the expense of the real-world safety of their platforms. When seeking to devise a pilot study around this question, it is advisable to consider how to avoid encouraging such behaviors.<sup>43</sup>

39 Artur. P. Carvalho et al., “Anonymisation and Compliance to Protection Data: Impacts and Challenges into Big Data,” ICEIS 1 (May 2021): 31–41.

40 Carvalho et al., “Anonymisation and Compliance.”

41 Nick Rasmussen, “GIFCT HRIA Response Letter,” GIFCT, November 19, 2021, <https://gifct.org/2021/07/20/hria-response-letter-by-nick-rasmus-sen/>.

42 Michael D. Ekstrand, Rezvan Joshaghani, and Hoda Mehrpouyan, “Privacy for All: Ensuring Fair and Equitable Privacy Protections,” in Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Proceedings of Machine Learning Research 81 (2018): 35–47, <https://proceedings.mlr.press/v81/ekstrand18a.html>.

43 For example, see Peter Bright, “Tumblr’s porn ban is going about as badly as expected,” ArsTechnica, December 5, 2018, <https://arstechnica.com/gaming/2018/12/tumblr-porn-ban-is-going-about-as-badly-as-expected/>; Louise Matsakis, “Tumblr’s Porn-Detecting AI has One Job – and it’s Bad at It,” Wired, December 5, 2018, <https://www.wired.com/story/tumblr-porn-ai-adult-content/>; Samantha Allen, “Why YouTube Wants to Hide These LGBT Videos From Young People,” Daily Beast, April 10, 2017, <https://www.thedailybeast.com/why-youtube-wants-to-hide-these-lgbt-videos-from-young-people/>; Maarten Sap et al., “The Risk of Racial Bias in Hate Speech Detection,” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (July, 2019), 1668–1678.

## Next Steps

Given the limitations described above, the ethical risks identified, and the practicalities of implementing such a project, gaining agreement to implement a methodology similar to the above is highly unlikely. Expressed at a high-level, a request for specific data with reasonable safeguards in place to mitigate privacy concerns becomes fraught with complex legal and technical challenges when the detail of the request is examined. What the specific data requested is matters a great deal to both the technical viability of any methodology and the legality of such a data disclosure. This methodology's request for specific raw data could utilize a company's pre-existing data gathering mechanisms for internal research, but further work is needed to address the scope of data requested before it could be confirmed that this practice would provide reasonable safeguards to mitigate the concerns around privacy and consent.

Furthermore, the safeguards put in place to mitigate privacy risks must be considered in full prior to such a request and include provisions such as a research code of conduct, research-ethics training for all people who access such data, guidance on how to collect consent and approach researching certain topics or user types, standardized consent forms and information sheets, an ethics expert on research review teams and data policies for user research.

The risks for the two different approaches to data disclosure outlined in the methodology are very different. Enhanced transparency reporting assumes a detailed publication process that sanitizes, summarizes, and aggregates data. Anonymized raw data disclosure is different in that it is inherently more intrusive to individual privacy and poses other challenges as described above.

Though the research question considered here is important, there is no viable route to enact the full scope of this methodology. In order to make progress, it must be redesigned to strengthen safeguards to privacy and ensure that the data requested is necessary and proportionate to the risks that are sought to be mitigated before further discussion with industry partners is appropriate.

In developing this methodology it was suggested that next steps should include the identification of specific GIFCT member companies willing to commit to taking this forward. However, the prevailing opinion of the TAWG was that given the level of ethical risk and the breadth of concerns identified, more work is needed in a multi-stakeholder forum before specific detailed plans for implementation can be considered. This work could take the form of focusing the methodology on enhanced transparency reporting rather than raw data disclosure.

GIFCT should identify a multi-stakeholder team to address the need for rescoping no later than the end of 2022.

## Overarching Considerations in Methodology Design

### Life, Liberty, and Security of Person

Ultimately, GIFCT's mission is to prevent terrorists and violent extremists from exploiting digital

services. This mission is grounded in the fact that everyone has the right to life, liberty, and security of person. As such studies must all contribute to an understanding of how terrorists make use of the internet, how the structure of the internet helps or hinders terrorists, and in particular when looking at how TVEC and borderline content is recommended the impact on the process of recruitment to terrorist or violent extremist groups and radicalization.

These methodologies must therefore not only concretely answer the appropriate research questions, but do so in a way that is necessary and proportionate to the threat faced in this situation.

## Scope

The scope of research into recommender systems and their impact on terrorism, violent extremism, and radicalization significantly impacts the design of methodologies and the considerations that need to be addressed to deliver effective, actionable, and responsible research. GIFCT member companies remove TVEC and so research restricted to this material will not elucidate the recommendation of other types of potentially harmful or radicalizing content with which the platform in question does not currently engage (since that content will ipso facto not be identified). Similarly, some companies also avoid recommending borderline content and depending on the definition adopted this could leave out a significant set of content. Beyond these two areas, research would extend into other online harms which have been linked to extremism such as hate speech<sup>44</sup>, misinformation, disinformation,<sup>45</sup> and conspiracy theories.<sup>46</sup> While there are undeniable links between these issues, as well as mainstream political speech, expanding the scope of the research as a result may have subsequent impacts in terms of privacy and the other considerations listed below.

Accordingly, each of the methodologies above seek to provide clarity on the prevalence and promotion of harmful content which tech companies have already identified as such and the success of their current mitigation strategies pertaining to it. It will not provide clarity on the current non-action of tech companies regarding other categories of potentially harmful/radicalizing content and hence the level of engagement that such content is allowed to garner. The latter is also of concern to policymakers; however, to expand the scope of this question to cover it presupposes both an agreed-upon third-party definition of what legal-but-harmful content should be covered and a library of corresponding content that can be used to train each platform's machine learning algorithms to accurately identify it, neither of which currently exist.

## Definitions

As we have detailed above, at best there are descriptions of TVEC and borderline content but no

.....  
 44 Florence Keen, "Banning Nazis or 'Burning Books'? How Big Tech is Responding to Hate Speech, and the Implications," GNET, July 2, 2020, <https://gnet-research.org/2020/07/02/banning-nazis-or-burning-books-how-big-tech-is-responding-to-hate-speech-and-the-implications/>.

45 Beatriz Buarque, "Why Some Far-Right Circles are Contributing to Vladimir Putin's Disinformation Campaign," GNET, March 21, 2022, <https://gnet-research.org/2022/03/21/why-some-far-right-circles-are-contributing-to-vladimir-putins-disinformation-campaign/>.

46 Elise Thomas, "Conspiracy Extremism and Digital Complexity – Where to Start?," GNET, October 5, 2020, <https://gnet-research.org/2020/10/05/conspiracy-extremism-and-digital-complexity-where-to-start/>; Marc-André Argentino and Amarnath Amarasingam, "The COVID Conspiracy Files," GNET, January 25, 2021, <https://gnet-research.org/2020/04/08/the-covid-conspiracy-files/>.

consistent definitions. Following GIFCT's work in 2021 on the taxonomy for the hash-sharing database, a need to explore further definition frameworks was identified, and GIFCT began work to build such a framework, analyzing definitions from 64 countries or intergovernmental organizations.<sup>47</sup> However, while this will help to standardize approaches to definitions and inform company policies, it will not bring a consensus across all parties.

A lack of standardization across definitions limits the generalizability of any research conducted, meaning that drawing conclusions across platforms or jurisdictions is unlikely to be achieved in the short term. The danger in drawing such conclusions is developing overly broad approaches to policy, legislation, and safeguards, which lead to unintended consequences in areas where the mechanics of how any effect operates are not well understood.

Furthermore, as Bharath Ganesh has highlighted in his 2021 article "Platform Racism," highly effective moderation by a tech company with a particular type of prohibited content may obfuscate the fact that they are using a narrow and minimal definition of this content.<sup>48</sup> Conversely, an overly broad definition may lead to the appearance that not enough is being done by a given platform.

How broadly "TVEC-adjacent content" is defined affects not only data sparsity but also the scope of lawful protected speech that is nevertheless being treated as suspect. There is a significant risk for bias and disproportionate scrutiny/impact to work its way into any given study based on the definition chosen and the recognition that each company uses their own definitions.

If we restrict the study to categories of content that companies are already banning, this may limit the concern. Balancing the academic interests in expanding scope to understand the full landscape in which these algorithms operate versus concerns around adverse human rights impact, but issues around generalizability and feasibility of meta-analyses remain.

## Impact on Terrorism

The research performed and the focus given to these challenges must be proportional to the threats from terrorists and violent extremists and balanced against the other research priorities in this area. We also need to consider the beneficial impacts of algorithms and that "whereas algorithms pose (un)known challenges for extremism, the opportunities they present in the mitigation and resolution of this and other societal challenges are equally consequential."<sup>49</sup> Priority should be given to research that is actionable, that can have a real impact on terrorism and violent extremism online, and that can show causality and agency so that interventions and policies can be driven by the evidence.

## User Privacy and Data Disclosure

Article 12 of the Universal Declaration of Human Rights says, "No one shall be subject to arbitrary

.....

<sup>47</sup> GIFCT Definitional Frameworks

<sup>48</sup> Bharath Ganesh, "Platform Racism: How Minimizing Racism Privileges Far Right Extremism," Social Science Research Council – Items, March 16, 2021, <https://items.ssrc.org/extremism-online/platform-racism-how-minimizing-racism-privileges-far-right-extremism>.

<sup>49</sup> Jazz Rowa - The Contextuality of Algorithms: A Human Security Approach to (Non)Violent Extremism in the Cyber-Physical Space - 2022



interference with his privacy, family, home, or correspondence. Everyone has a right to the protection of the law against such interference or attacks.”<sup>50</sup> This right needs to be balanced against the need to disclose information. The tradeoffs at play here are complex, multifaceted, and very much need to be assessed on a case-by-case basis. Some of the key considerations as laid out by Daphne Keller in her blog post “User privacy vs. platform transparency”<sup>51</sup> include:

- Who gets access
- How data is used
- How to manage content that discloses personally identifiable information
- How to manage data shared privately
- Data aggregation and anonymization
- Longitudinal studies

Each of the methodologies presented above has to address these issues to a greater or lesser degree, and depending on the platform the calculus for each will be different as users operate differently on different platforms, different expectations of privacy exist, and different terms of service and community guidelines apply.

W3C has recently published a set of privacy principles that should guide the development of the Web as a trustworthy platform as part of the Technical Architecture Group.<sup>52</sup> These principles should be used to help guide future development and improvement of methodologies addressing content-sharing algorithms and radicalization.

## Who Gets Access to Data?

In each of the proposed methodologies, decisions must be made about who qualifies to get access to the data as well as how they are trained and vetted. In most cases, data must be accessed by a trained researcher/NGO or a vetted government agency operating in the context of their work within an appropriate code of conduct.

However, as noted in the third methodology that we considered, there are reasons for a less regulated, more public release of data, providing a greater impetus for tech companies to address any evidence of TVEC content and borderline content being algorithmically recommended on their platforms. Similarly limiting data access to academics restricts groups such as journalists and other parts of civil society, who provide significant contributions to data and research in this area.<sup>53</sup> Conversely, as suggested in the second methodology, less direct access to data means that research

50 Universal Declaration of Human Rights, United Nations, December 10, 1948, <https://www.un.org/en/about-us/universal-declaration-of-human-rights>.

51 Daphne Keller, “User Privacy vs. Platform Transparency: The Conflicts Are Real and We Need to Talk About Them,” Center for Internet and Society, April 6, 2022, <https://cyberlaw.stanford.edu/blog/2022/04/user-privacy-vs-platform-transparency-conflicts-are-real-and-we-need-talk-about-them-0>.

52 “Privacy Principles,” World Wide Web Consortium (W3C), May 12, 2022, <https://www.w3.org/TR/2022/DNOTE-privacy-principles-20220512/>.

53 Keller, “User Privacy vs. Platform Transparency.”

can be carried out without compromising proprietary information, intellectual property, or user privacy.

In an area where trust must be established between sectors to effectively progress the collective understanding and inform effective policy and design, researcher independence is a key consideration. Being overly prescriptive about the requirements for researchers can limit this independence.

In general striking the balance among open access, trust, and protection of privacy and information should be addressed in a pragmatic manner and work should be undertaken to help provide clarity about what standard best practices should be and what qualifies as research to gain access to data building on the information provided in Appendix C.

## Precision and Recall

When assessing content moderation we must consider both “recall” metrics and “precision” metrics. Recall is the extent to which we can select all of the relevant posts in the dataset without leaving any out (false negatives). Precision is the rate at which from our dataset of posts we can select the relevant posts (true positives) without also getting any irrelevant posts (false positives).<sup>54</sup>

As noted early in the paper, recommender systems can be considered as a form of content moderation. To understand their functioning, some methodologies will focus on one or other of these metrics, but to appreciate the full picture and performance of the system we need to be able to understand both. A report which indicates near comprehensive moderation of in-scope content will not reveal if this has been achieved at the expense of moderating a high proportion of innocuous speech as well. There are studies (such as that by Dinar) that have shown that downranking can disproportionately impact vulnerable groups, and so it is essential that research is explored with respect to both aspects to avoid drawing conclusions that inform policies, safeguards, and interventions that inadvertently have adverse impacts on these groups.<sup>55</sup>

## Security Safeguards

Data disclosed to appropriately qualified researchers as part of a well-designed and responsible study must also be protected to ensure that the data is not lost and that user privacy and security is not compromised. Criteria, standards, and best practices for privacy, security, and confidentiality must be in place before data can be shared. Before engaging in research projects there is a duty on both researchers and tech platforms to ensure that the systems in place provide reasonable mitigation to cyber security risks. Data handling and security procedures must also be in compliance with regulations such as the GDPR. However, in developing these standards, care must be taken to ensure

.....

54 Thorley, T. & Saltman, E. (2022, June 28 - 29). GIFCT Tech Trials: Combining Behavioural Signals to Surface Terrorist Content Online, [Conference Presentation]. Terrorism and Social Media Conference, Swansea University, Wales. <https://www.tasmconf.com/>.

55 Christina Dinar, “The state of content moderation for the LGBTIQ+ community and the role of the EU Digital Services Act,” Heinrich-Böll-Stiftung, European Union, June 21, 2021, <https://eu.boell.org/en/2021/06/21/state-content-moderation-lgbtiga-community-and-role-eu-digital-services-act>.

that the cost of implementation does not preclude the research and prevent its viability.

## Research Codes of Conduct

Companies' internal processes for studying the effects of different recommender system experiences may also be subject to internal ethical controls that should be considered before preparing a pilot study methodology. For instance, companies may adopt a code of conduct or guidelines with detailed information on topics such as collecting consent and how to approach researching certain topics or user types. They may similarly use standardized consent forms and information sheets that would allow them to follow a template for each study with consistent language.

Such controls are implemented to address ethical concerns across research projects and should be understood and thoughtfully considered before pilots are developed. Such protections can help to ensure that pilots adequately assess the range of ethical concerns that may be present on the platform. Further, if the pilot study conflicts with the company's own ethical compliance protocol, this may significantly impact the ability to conduct and release research data without violating existing company protections (or even applicable law).

## Equality and Non-discrimination

All human beings are born free and equal in dignity and rights, and everyone is entitled to all rights and freedoms without distinction of any kind such as race, color, sex, language, religion, political or other opinion, national or social origin, property, birth, or status. Limitations of research design in this space require either very broad datasets to cover the vast range of different groups that may be impacted by algorithms and ensure that biases can be identified that have implications for privacy or targeted studies that may disproportionately impact vulnerable groups and not highlight biases or not be generalizable.

## Conclusion

The role of the internet in individuals becoming radicalized to join violent extremist groups or commit violent acts motivated by extremist ideologies has been well documented. The process of radicalization, the subsequent harm and horrific attacks that can occur, and the online aspects of terrorism and violent extremism must be fully understood and addressed. As a result, GIFCT seeks to prevent terrorists and violent extremists from exploiting digital platforms. GIFCT member companies seek to remove TVEC and have made significant improvements in how they manage content-sharing algorithms in order to mitigate potential risks. Safety by design is a core part of this process and assurance is needed that when adding a feature or technology to the web, the harm it could do to society or groups (especially to vulnerable people) has been considered and where possible mitigated.

While we aimed to reach consensus, this paper highlights the debates and counterpoints to various issues where a consensus position has not been reached. In this paper, we have identified key research questions that still need to be resolved and produced a taxonomy to help ensure that gaps

in the research can be identified and addressed. We then focus on three of these key questions, evaluating different methodologies and data disclosure processes to address them. In doing so we have identified several key areas that need to be addressed in designing studies in this field and practical ways forward to navigate the nuances in this field with a responsible and human rights-based approach.

We conclude that to properly address technical approaches that answer these research questions, methodological design must address definitional issues, generalization, privacy and security, a range of human rights, and ultimately the impact on terrorism and violent extremism. Further, it is vital that tech companies both engage and are engaged in the design process and assessment of methodologies as they have the knowledge and expertise to understand what is feasible and what data and infrastructure is available. Pilot studies such as discussed in the methodologies above provide a concrete and practical focus for this engagement and allow companies to evaluate specific issues and iterate towards an appropriate, responsible, and impactful solution.

This research also requires significant resources to conduct, and models for funding and ensuring capacity in the research community to address this and other issues at the intersection of terrorism and technology (such as those employed by GPAI and GNET) should be supported.

## Recommendations

In writing this paper we aimed to both seek consensus between the multistakeholder participants of the GIFCT TAWG and highlight the debates and counterpoints to various issues where a consensus position has not been reached.

The methodologies and pilot studies discussed in this paper should not be considered as a commitment to conduct the pilot studies but a commitment to discuss the feasibility of the methodology and how they could be taken forward or redesigned. This is a continuing effort and will be an iterative process.

This paper evaluates the feasibility of three proposed pilot study methodologies for researching the intersection of content recommender systems and radicalization, identifying issues that prevent studies using these methodologies from moving forward, and next steps to take in iterating the research design.

## Recommendations for GIFCT

### Question 1: What users are most likely to have borderline content recommended to them?

- GIFCT should seek to identify a research team with the capacity to further the design and implementation of this project no later than October 2022.
- GIFCT should seek to arrange meetings between specific GIFCT member companies (including relevant technical experts) and the research team to explore the technical viability of this project, with a view to reaching a decision no later than the end of 2022.

## Question 2: What are the effects of recommender systems on platform users' attitudes towards TEVC?

- GIFCT should seek to arrange meetings between specific GIFCT member companies and the research team to discuss technical aspects of this project (for instance, appropriate metrics for measuring users' attitudes towards TVEC), with a view to reaching a decision no later than the end of 2022.
- GIFCT should seek to arrange meetings between specific GIFCT member companies and the research team to discuss legal aspects of this project (relating to privacy and consent), with a view to reaching a decision no later than the end of 2022.

## Question 3: How is TVEC and borderline content that is ultimately moderated recommended by content-sharing recommender systems before and after moderation takes place?

- This methodology should be rescoped and redesigned to strengthen safeguards to privacy and ensure that the data requested is necessary and proportionate to the risks that are sought to be mitigated, perhaps focusing on enhancing transparency reporting as the disclosure method rather than raw data publication.
- GIFCT should identify a multi-stakeholder team to address the need for rescoping no later than October 2022.

## Recommendations for Tech Companies

- Each of the research questions that were identified (Appendix A) in this process represents gaps in knowledge about the intersection of users and content and the potential implications for radicalization. Tech companies could help with research and policy to address these gaps in knowledge by comparing the research questions with existing evaluations of their platforms and content moderation practices. Where existing work does not address the research questions, tech companies could suggest feasible methodologies to study these areas.
- Evaluation of the methodologies explored in this paper was a significant undertaking given the complexity of the internal processes, expertise, and teams needed to be consulted within tech companies. There is, and will continue to be, a focus on third-party or independent research into these research questions. Developing processes and identifying efficiencies in evaluating research proposals would make a significant difference in answering these research questions.

## Recommendations for Researchers and Policy Makers

- An understanding of the relative impact and causal mechanisms at play is critical to mitigating risks in this space. However, as it has been noted, "the internet's worst websites aren't algorithmic."<sup>56</sup> The investment in this research should be proportionate to the impact on terrorist and violent extremist activity online, be conducted in a human rights-based manner, and be prioritized holistically against

.....  
 56 Ryan Broderick, "You can't always blame algorithms," May 16, 2022, <https://www.garbageaday.com/p/you-cant-always-blame-algorithms>.

other research areas aimed at preventing terrorists and violent extremists from exploiting digital platforms.

- Gaps remain in understanding how recommender algorithms operate. Though much has been said publicly, a systematic meta-analysis of what is being disclosed already by companies as well as a thorough gap analysis assessing currently available information is called for.
- Beyond the scope of this paper, but core to the question of how much agency recommender systems have in this process, is understanding how borderline content impacts users and user behavior with regard to radicalization and progression to terrorism or violent extremism. Existing research should be reviewed and the gaps identified should be used to commission further work.
- Safeguards, policies, and positive interventions to mitigate risks of recommender systems contributing to radicalization should be considered and designed to inform pilot studies and methodological design aimed at answering the identified research questions (Appendix A). However, implementation of such interventions would be premature without a solid understanding of the causal mechanisms at play.
- Cultivating more independent researchers to identify methodologies and propose pilot studies.

## Further Reading

- Ada Lovelace Institute. Inspecting algorithms in social media platforms. November, 2020. <https://www.adalovelaceinstitute.org/wp-content/uploads/2020/11/Inspecting-algorithms-in-social-media-platforms.pdf>.
- “Content personalisation and the online dissemination of terrorist and violent extremist content,” Tech Against Terrorism. February, 2021. <https://www.techagainstterrorism.org/wp-content/uploads/2021/02/TAT-Position-Paper-content-personalisation-and-online-dissemination-of-terrorist-content1.pdf>.
- Criezis, Meili. “Remaining and Expanding or Surviving and Adapting? Extremist Platform Migration and Adaptation Strategies.” GNET. November 12, 2021. <https://gnet-research.org/2021/11/12/remaining-and-expanding-or-surviving-and-adapting-extremist-platform-migration-and-adaptation-strategies/>.
- Bakshy, Eytan, Messing, Solomon, and Adamic, Lada A. “Exposure to ideologically diverse news and opinion on Facebook.” Science 348, no. 6239 (May, 2015): 1130–1132. <https://doi.org/10.1126/science.1241160>.
- Decker, Benjamin T., and Boucher, Tim. “Disrupting Online Harms: A New Approach.” Global Disinformation Index. July 23, 2021. <https://disinformationindex.org/wp-content/uploads/2021/07/2021-07-23-Disrupting-Online-Harms-A-New-Approach.pdf>.
- Frissen, Thomas. “Internet, the great radicalizer? Exploring relationships between seeking for online extremist materials and cognitive radicalization in young adults.” Computers in Human Behavior 114 (January 2021): 106549. <https://doi.org/10.1016/j.chb.2020.106549>.
- Keller, Daphne. “Amplification and Its Discontents.” Knight First Amendment Institute. June 8, 2021. <https://knightcolumbia.org/content/amplification-and-its-discontents>.
- Kfir, Isaac. “Algorithms, the Search for Transcendence and Online Radicalisation.” GNET. October

- 14, 2021. <https://gnet-research.org/2021/10/14/algorithms-the-search-for-transcendence-and-online-radicalisation/>.
- Ledwich, Mark, and Zaitsev, Anna. "Algorithmic extremism: Examining YouTube's rabbit hole of radicalization." First Monday. March 2, 2020. <https://doi.org/10.5210/fm.v25i3.10419>.
  - O'Connor, Ciarán. "Hatescape: An In-Depth Analysis of Extremism and Hate Speech on TikTok." Institute for Strategic Dialogue. November 29, 2021. <https://www.isdglobal.org/isd-publications/hatescape-an-in-depth-analysis-of-extremism-and-hate-speech-on-tiktok/>.
  - Rose, Hannah, and C., A. "Youth-on-Youth Extreme-Right Recruitment on Mainstream Social Media Platforms." GNET. January 10, 2022. <https://gnet-research.org/2022/01/10/youth-on-youth-extreme-right-recruitment-on-mainstream-social-media-platforms/>.
  - Rowa, Yvonne Jazz. "Part 1: Algorithmic Deconstruction in the Context of Online Extremism." GNET. September 15, 2020. <https://gnet-research.org/2020/09/15/part-1-algorithmic-deconstruction-in-the-context-of-online-extremism/>.
  - Rowa, Yvonne Jazz. "Part 2: Algorithmic Agency in Online Extremism: The Bigger Picture." GNET. September 21, 2020. <https://gnet-research.org/2020/09/21/part-2-algorithmic-agency-in-online-extremism-the-bigger-picture/>.
  - Thomas, Elise. "Recommended Reading: Amazon's algorithms, conspiracy theories and extremist literature." Institute for Strategic Dialogue. November 23, 2021. <https://www.isdglobal.org/isd-publications/recommended-reading-amazons-algorithms-conspiracy-theories-and-extremist-literature/>.
  - Wolfowicz, Michael. "Examining the interactive effects of the filter bubble and the echo chamber on radicalization." Journal of Experimental Criminology (August 3, 2021). <https://link.springer.com/article/10.1007/s11292-021-09471-0>.
  - "YouTube Regrets." Mozilla Foundation. July, 2021. <https://foundation.mozilla.org/en/youtube/findings/>.



## Appendix A: Full List of Research Questions Considered

### 1. What are the characteristics of users that increase the chances that they will be recommended borderline content?

- a. **Selected question:** What users are most likely to have borderline content recommended to them?
- b. What users are most vulnerable to being suggested terrorist or violent extremist (or "borderline") content?
- c. What user behaviors prompt exposure to recommendations for borderline content?
- d. What are the differences between groups being provided different approaches to surfacing content (e.g., recommendations versus no recommendations or different versions of recommender algorithms)?
- e. How are illegal terms and conditions (T&C)-breaching content, legal but T&C-breaching content, and legal borderline content present on online platforms broken down in terms of type (e.g., hate speech, mis/disinformation, TVEC) and distribution (demographics of ages, geographical location, etc.)?

### 2. What are the characteristics of borderline content that increases chances that it will be recommended to users?

- a. What is the relative reach of TVEC versus borderline versus innocuous content?
- b. Is there a difference between the rate at which innocuous content is recommended versus borderline versus TVEC?
- c. What percentage reach of borderline content (and/or TVEC) is the result of the content being recommended and is this different compared to innocuous content?
- d. What is the poster-consumer relationship for illegal T&C-breaching content, legal but T&C-breaching content, and legal borderline content consumed on online platforms?
- e. What percentage of consumption is the result of the content being algorithmically promoted to newsfeeds / recommended content lists/search results?
- f. Does the poster have a history of posting/sharing such content?
- g. How is consumption related to consumers' relationships to the poster/sharer? (What percentage of consumers follow the poster? What percentage consumed it as public content?)
- h. What proportion of consumers have a history of consuming this type of content?

### 3. What is the impact of Content Recommending System on Users' Behavior?

- a. **Selected question:** What are the effects of recommender systems on platform users' attitudes towards TVEC content?
- b. Is there a difference in the engagement of users with recommended borderline versus non-recommended content?
- c. What features or functions of recommender systems have the greatest impact on driving people toward (or away from) violent extremism?

- d. What are the ways in which recommended systems reinforce or dispel extremist views held in particular groups or communities?
- e. How to possibly assess the risk of radicalization on a platform (or some parts of it)? Can we identify causal links between the use of algorithms and potential radicalization, and based on what data and factors?
- f. Is it possible to assess the degree to which an algorithm is more or less capable to lead to radicalization based on observing its behavior (e.g., across users), if possible?

#### 4. What is the impact of borderline content on users?

#### 5. What is the impact of Content Recommending System on the reach of borderline content?

- a. How do online platforms' open engagement-driven recommender algorithms interact with borderline and T&C-breaching content?
- b. Selected Question: How is TVEC content that is removed recommended by content-sharing recommender systems before removal takes place?
  - i. How many users has it been recommended to?
  - ii. How many users have consumed it?
  - iii. What is the relative reach of TVEC that has been recommended prior to removal versus TVEC that has not?
- c. What is the promotion journey of illegal T&C-breaching content, legal but T&C-breaching content, and legal borderline content?
  - i. How has the content been promoted and consumed over 7 days, 30 days, 90 days, etc., until it is moderated?
  - ii. What is the average viewership, time delay, and reach of moderated content before it is moderated?
  - iii. What form did moderation take? Where moderation comes in the form of "downranking," how did that affect the subsequent promotion and consumption over 7 days, 30 days, and 90 days?
  - iv. What proportion of subsequently moderated content was initially promoted by recommender algorithms?
- d. What factors may affect whether (and if so) to what degree algorithms can amplify TVEC dissemination and radicalization?

#### 6. What characteristics of users are most likely to consume and share borderline content?

#### 7. Other questions considered:

- a. What mitigations are available to manage the risks of increased radicalization that recommender systems may pose and which are most effective at minimizing these risks?
- b. How can we audit and (perhaps most importantly) monitor the algorithms used to recommend content in order to ensure their beneficial/safe behavior?
- c. What processes and tools may be needed (by platforms/trusted flaggers/LEAs, etc.) to manage any risks created by these algorithms?

## Appendix B: Vulnerable Groups

We should pay particular attention to the rights, needs, and challenges of individuals from groups or populations that may be at heightened risk of becoming vulnerable. Vulnerable groups are those that face being marginalized, discriminated against, or exposed to other adverse human rights impacts with greater severity and/or lesser potential for remediation than others.

Vulnerability depends on context, and someone who may be powerful in one context may be vulnerable in another. Examples include:

- **Formal Discrimination:** Laws or policies that favor one group over another.
- **Societal Discrimination:** Cultural or social practices that marginalize some and favor others.
- **Practical Discrimination:** Marginalization due to life circumstances, such as poverty.
- **Hidden Groups:** People who might need to remain hidden and consequently may not speak up for their rights, such as undocumented migrants.

Though every case is unique, here are examples of vulnerable groups:

|                                                              |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
|--------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Aboriginal/<br>Indigenous<br>peoples                         | Aboriginal or indigenous peoples have a historical existence and identity that is separate and independent of the states now enveloping them on account of their descent from populations that inhabited the geographical region to which the country belongs at the time of colonization or establishment of present state boundaries. This group, irrespective of their legal status, retains some or all of their own social, economic, cultural, and political institutions.                                                                      |
| Age-related<br>groups                                        | Groups of specific age, particularly the young or very old, that experience particular vulnerabilities, such as medical or social exclusion or discrimination.                                                                                                                                                                                                                                                                                                                                                                                        |
| Disability                                                   | Any condition of the body or mind that makes it more difficult for the person with the condition to do certain activities (activity limitation) and interact with the world around them (participation restrictions). This includes people who have a record of such an impairment, even if they do not currently have a disability. It also includes individuals who do not have a disability but are regarded as having a disability. Discrimination against this group may also include those with an association with a person with a disability. |
| Historically<br>oppressed<br>ethnic or racial<br>communities | Social groups that have a common national or cultural practice, tradition, and perspectives, or shared physical or social qualities that are viewed as distinct by society that have been subject to harsh and authoritarian treatment.                                                                                                                                                                                                                                                                                                               |
| Non-binary<br>gender identity                                | Persons that fall within a spectrum of gender identities that are not exclusively masculine or feminine.                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| Homeless /<br>Underhoused                                    | Persons who lack a fixed, regular, and adequate nighttime residence or that sleep in a shelter designated for temporary living accommodations or in places not designated for human habitation.                                                                                                                                                                                                                                                                                                                                                       |

|                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|----------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Immigrants, refugees, and migrants     | Persons legally or illegally outside of their country of usual residence. This group also includes refugees, who are outside their country of origin for reasons of feared persecution, conflict, generalized violence, or other circumstances that have seriously disturbed public order and, as a result, require international protection.                                                                                                                                 |
| Incarcerated people and their families | Groups of people who either have been imprisoned or have familiar ties with individuals who have been imprisoned.                                                                                                                                                                                                                                                                                                                                                             |
| Linguistic communities                 | A community that shares a set of linguistic norms and speech.                                                                                                                                                                                                                                                                                                                                                                                                                 |
| Low-income people or communities       | Persons that do not meet income state requirements to be considered middle-class and may be struggling with financial insecurity.                                                                                                                                                                                                                                                                                                                                             |
| Faith or belief-based communities      | Persons whose values are based on faith and/or beliefs, and which most often draws its activists (e.g., leaders, staff, volunteers) from a particular faith group, including but not limited to types of Christianity, Hinduism, Islam, Judaism, Sikhism, Buddhism, and Baha'i, including minorities and dissenters within those communities, as well as persons who have renounced or changed their faith, as well as communities who define as atheistic (e.g., humanists). |
| Inner-urban communities                | Communities located in central areas of cities that may experience social and economic disparity relative to the rest of the surrounding area or city.                                                                                                                                                                                                                                                                                                                        |
| Rural communities                      | Populations residing in rural areas or countryside located outside towns and cities that may experience varied rates of poverty, unemployment, insurance, and access to education and health compared to their urban counterparts.                                                                                                                                                                                                                                            |
| LGBTQI+                                | Persons who identify as lesbian, gay, bisexual, transgender, queer, intersex, and others.                                                                                                                                                                                                                                                                                                                                                                                     |
| Human rights defenders                 | Persons who, individually or with others, act to promote or protect human rights, such as human rights organizations, journalists, citizen journalists, political activists, and members of other vulnerable groups advocating for their rights. Human rights defenders are identified above all by what they do, and it is through a description of their actions and of some of the contexts in which they work that the term can best be understood.                       |
| Caste                                  | Hereditary social classes that restrict the occupation of their members and their association with the members of other castes; a system of rigid social stratification characterized by hereditary status, endogamy, and social barriers sanctioned by custom, law, or religion.                                                                                                                                                                                             |

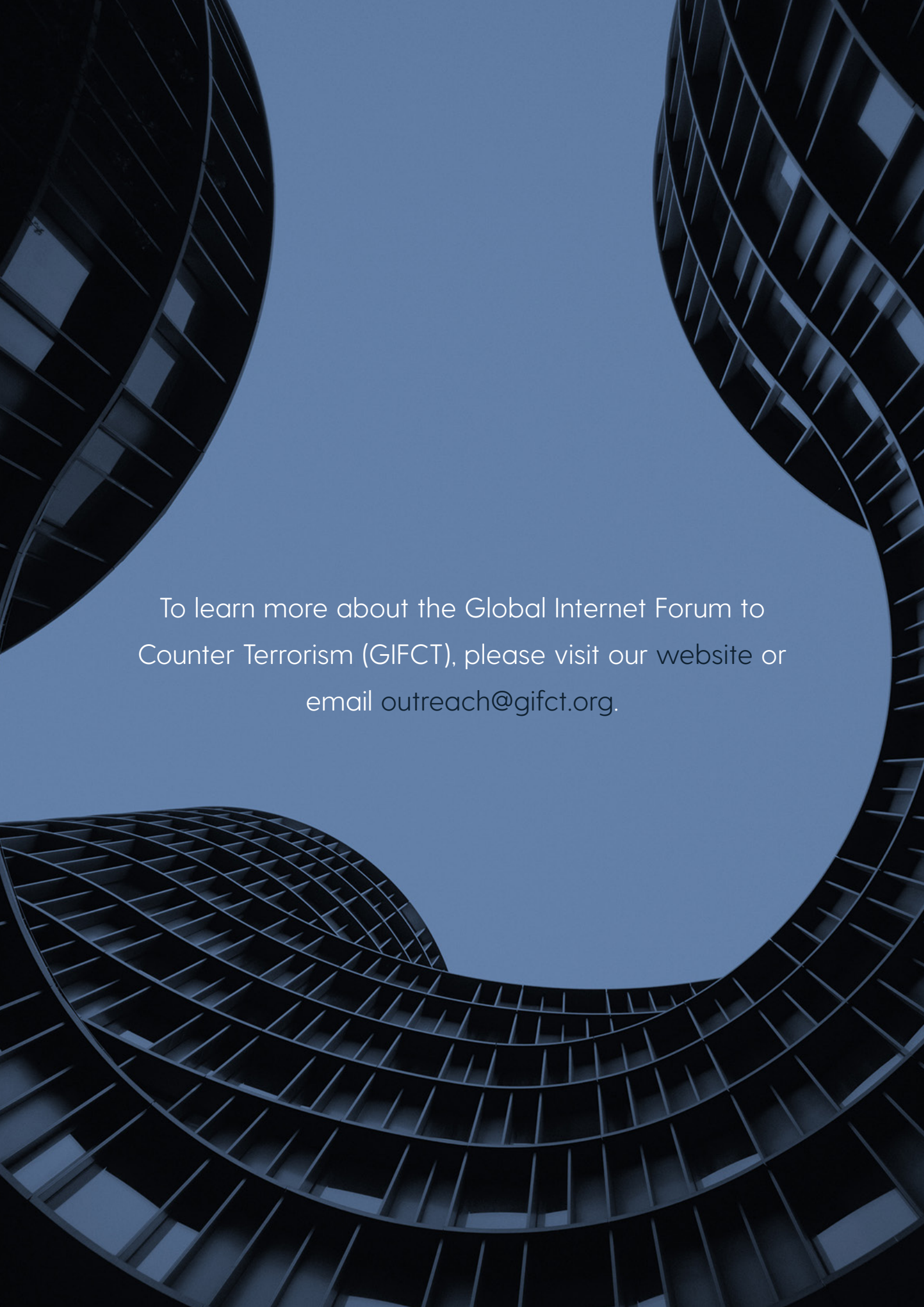
## Appendix C: Tech Platform Research Review Considerations

Tech Platforms are likely to have some/all of the following which guide their engagement in research pilots:

- **A code of conduct:** Some organizations write their own code of conduct so that it is as relevant as possible to their user-research context. Other organizations might adopt a professional body's code of conduct. They may cite adherence to the code of conduct in participant communication (e.g., Google uses APA<sup>57</sup>).
- **Research-ethics training for all people who carry out user research:** This type of training may be included in onboarding, e-learning, or ad-hoc training courses.
- **Guidance documents:** Organizations often have guidelines on how to collect consent, how to write good consent forms and information sheets, and how to approach researching certain topics or user types.
- **Standardized consent forms and information sheets:** Mature organizations have standardized study documents which contain areas where researchers can fill in the details about the study while keeping the core language consistent.
- **Ethics experts:** These could be people on a review team or service providers who deliver training, provide advice, or share knowledge with the team.
- **Data policies for user research:** Organizations have a specific data policy for UX teams carrying out user research; this policy covers relevant data protection laws and how the organization complies with them. It includes what constitutes personal data produced during or from user-research activities, where it gets stored, and how it is handled.

.....  
 57 "Ethical Principles of Psychologists and Code of Conduct," American Psychological Association, January 1, 2017, <https://www.apa.org/ethics/code>.





To learn more about the Global Internet Forum to  
Counter Terrorism (GIFCT), please visit our website or  
email [outreach@gifct.org](mailto:outreach@gifct.org).