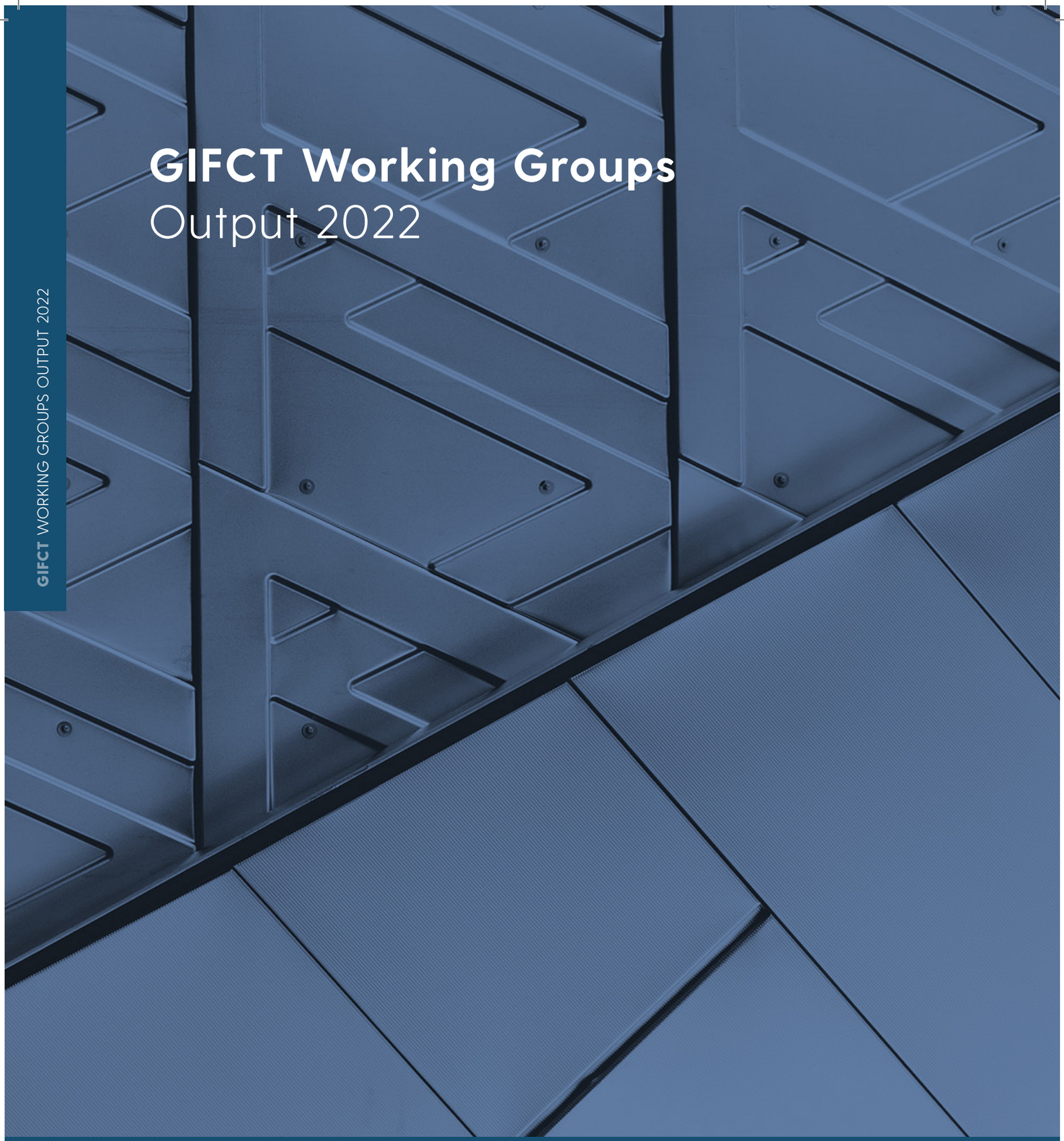
GIFCT Working Groups
Output 2022

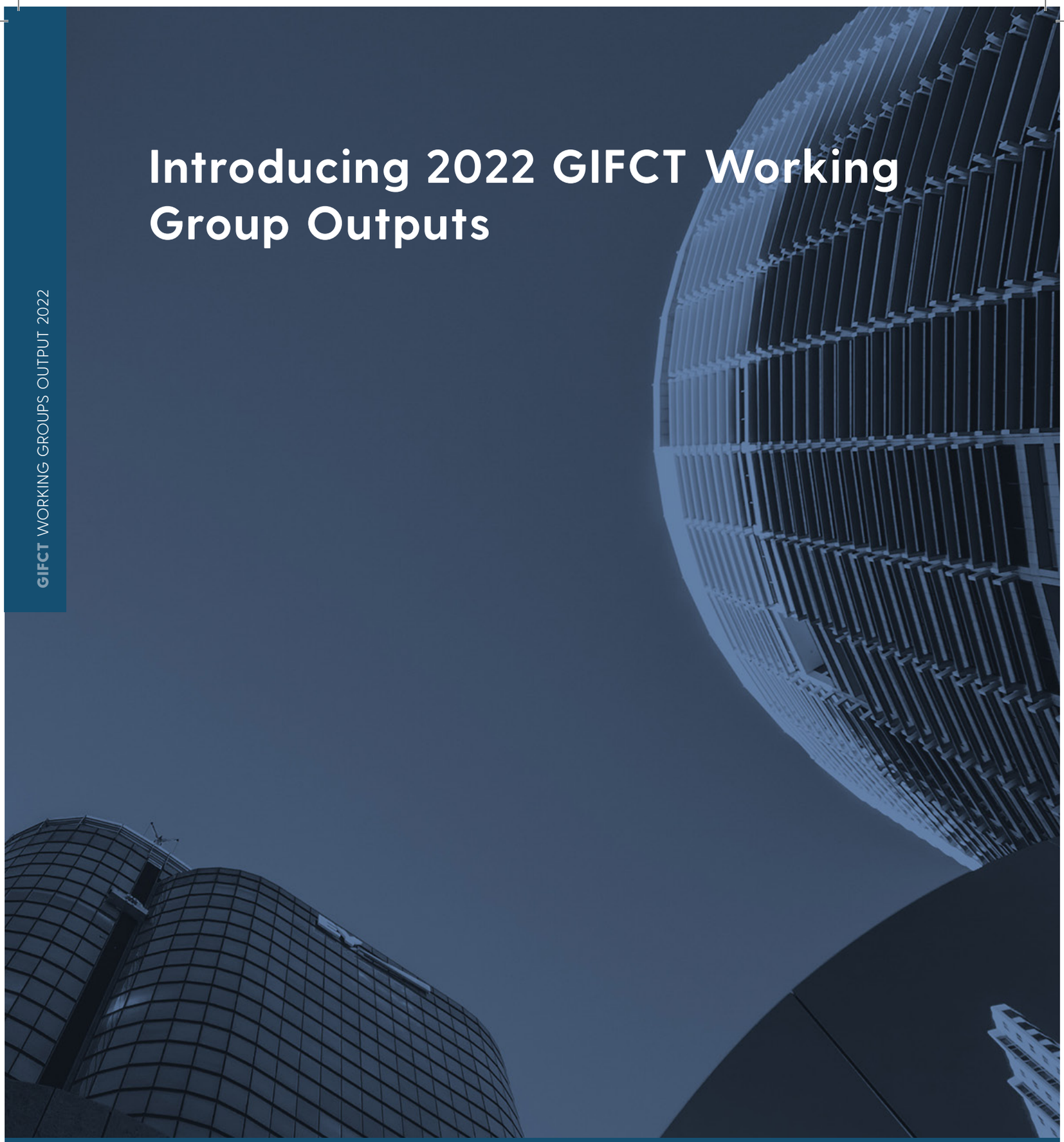


GIFCT
Global Internet Forum to Counter Terrorism

Table of Contents

GIFCT WORKING GROUPS OUTPUT 2022

Introducing 2022 GIFCT Working Group Outputs



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 Vaughn, 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:

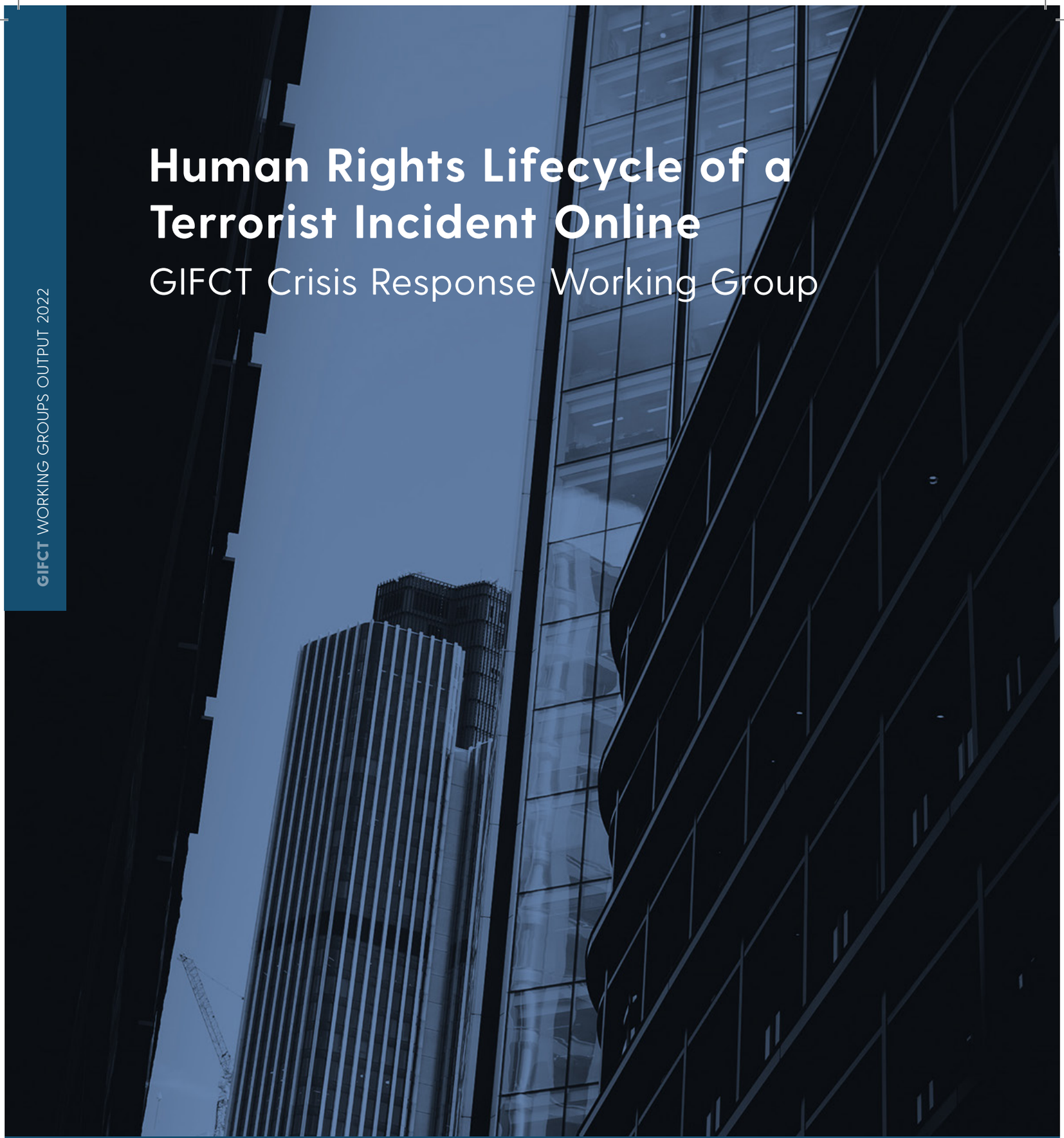
Tech Sector	Government Sector	Civil Society / Academia / Practitioners	Civil Society / Academia / Practitioners
ActiveFence	Aqaba Process	Access Now	Lowy Institute
Amazon	Association Rwandaise de Défense des Droits de l'Homme	Anti-Defamation League (ADL)	M&C Saatchi World Services Partner
Automattic	Australian Government - Department of Home Affairs	American University	Mnemonic
Checkstep Ltd.	BMI Germany	ARTICLE 19	Moonshot
Dailymotion	Canadian Government	Australian Muslim Advocacy Network (AMAN)	Moduslzad - 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 (1267 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	

GIFCT WORKING GROUPS OUTPUT 2022

Human Rights Lifecycle of a Terrorist Incident Online

GIFCT Crisis Response Working Group



GIFCT
Global Internet Forum to Counter Terrorism

Dr. Farzaneh Badii
Digital Medusa

Executive Summary

From January 15 to May 31, 2022, the Working Group on Crisis Response Protocols (CRWG) – a subgroup of the Global Internet Forum to Counter Terrorism (GIFCT) – met with stakeholders across civil society organizations, governments, academics and companies (through a series of individual and group meetings in addition to tabletop exercises). The output of this effort is the present report that aims to:

1. Outline the lifecycle of a terrorist incident on the Internet and its human rights impact;
2. Propose a framework for crisis protocol operators and GIFCT to use for explicating the lifecycle of incidents and to consider human rights implications in crisis response; and
3. Clarify the relationship between human rights and GIFCT's mission through explaining the human rights impact at each stage of the crisis lifecycle.

Broadly speaking, the output of the report is centered on an analysis of the lifecycle of a terrorist attack online, mapped against nine actual case studies (from Halle to Christchurch), and brings together the stages of the crisis protocol with their implications for human rights.

Crisis Protocol Stages: This section outlines the different stages of a terrorist incident on the Internet, from Horizon (before an attack takes place) to the Conclusion (which includes actions from standing down a response through to conducting debriefs). Definitions of each stage and particular categories of mapping, such as the type of attack or the virality of the attack, are available throughout the report.

Human Rights Principles: At each stage of a crisis, a number of potential human rights are potentially impacted. These human rights, which may include privacy, nondiscrimination and equality before the law, and access to effective remedy are mapped against not only the Crisis Protocol Stage but also the rightsholders based on a 2021 Human Rights Assessment undertaken by Business for Social Responsibility for GIFCT. Rightsholders include but are not limited to victims of terrorism and violent extremism, victims of efforts to counter terrorism and violent extremism, human rights defenders, the accused, and particularly groups spanning women, girls, and families as well as men and boys.

Using actual case studies coupled with a series of tabletop exercises enabled us to refine the proposed framework for crisis protocol operators in the event of a terrorist incident on the Internet. However, this framework should be considered a starting point for future discussions and investigations on how terrorism manifests online, appropriate and effective methods for crisis response across relevant groups, and the human rights potentially impacted at each stage of response. In collaboration with civil society, government, academic and industry partners, we look forward to continuing to grow and refine this framework in the future.

Background

The Working Group on Crisis Response Protocols (CRWG), a subgroup of Global Internet Forum to Counter Terrorism (GIFCT), drafted this report to:

1. Outline the lifecycle of a terrorist incident on the Internet and its human rights impact;
2. Propose a framework for crisis protocol operators and GIFCT to use for explicating the lifecycle of incidents and to consider human rights implications in crisis response; and
3. Clarify the relationship between human rights and GIFCT's mission through explaining the human rights impact at each stage of the crisis lifecycle.

This report also contributes to the Christchurch Call work plan for crisis response, which includes establishing due process and human rights protections, to ensure all protocols are developed and implemented in a robust way.

Method

We have used a mixed method approach to draft this report, but mainly employed an iterative process by interviewing several stakeholders involved with tackling online crises. These include civil society organizations, academics, tech-corporations and law enforcement agencies. We have also made use of the table-top exercises that were held by GIFCT to develop the indicators of human rights impact.

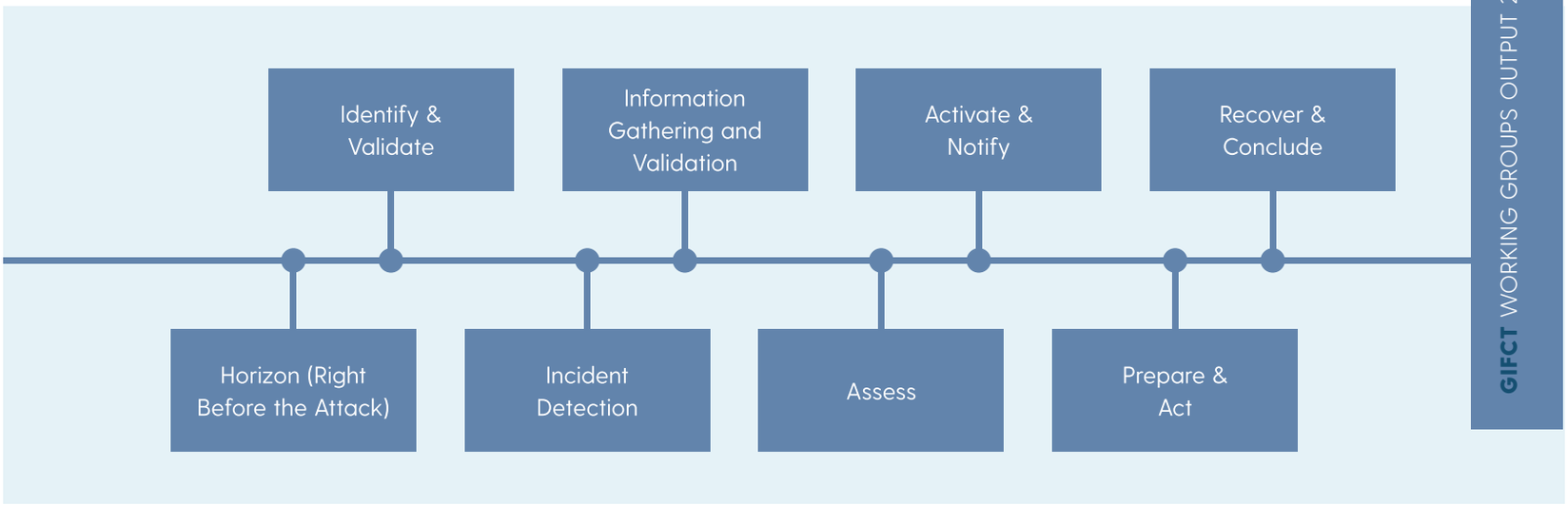
This report provides categorical tables about various terrorist and violent extremist incidents with an online impact. Through case studies it then describes each stage of the crisis protocol and its human rights impact and provides a framework for understanding the possible level of human rights impact at each stage of the crisis protocol.

Crisis Protocol Stages

The stages of crisis protocol described here are a combination of common practices of various crisis protocols and the sub-group's ideas. The horizon stage was specifically added by the sub-group. The stages are as follows:

1. Horizon (right before the attack)
2. Identify and validate (Stages 2 through 6 may involve contact/cooperation with third party governments and OSPs and industry bodies like Tech Against Terrorism (TAT)).
3. Incident detection (through internal monitoring, or a tip received from a partner organization, or media reporting).
4. Information gathering and validation/part of pre-activation (seeking information to understand what has happened or is happening and ensure that understanding is valid, i.e. corresponds to reality).
5. Assess/also part of pre-activation (whether the incident meets criteria/thresholds for activation, such as murder or mass violence, has a terrorist or violent extremist motivation, content was produced by perpetrator, accomplice or supporter, has potential to go viral).
6. Activate and notify (activate the protocol, notify the members, inform them of the level of action needed (monitoring or doing more), maybe also notify civil society organizations and the public).

7. Prepare and Act/active response and information sharing (look at Open Source Intelligence (OSINT) materials, share hashes and awareness about where the content is, take action to find/moderate/remove content, preserve data, share actions and outcomes, ongoing strategic communications).
8. Conclude (assessment against threshold, stand down response, notify members/stakeholders/public, may continue to monitor, documenting decisions/actions, organizing debrief/multistakeholder review, sharing findings with stakeholders and public).



Human Rights Principles

The human rights principles referred to above are based on the Business for Social Responsibility (BSR) human rights impact assessment of GIFCT.[1] They include:

1. Life, liberty, and security of person (UDHR 3;ICCPR 6, 9)
2. Nondiscrimination and equality before the law (UDHR 1, 2, 7; ICCPR 2, 3, 26; ICESCR 2, 3; CEDAW 2; CERD 2)
3. Access to effective remedy (UDHR 8; ICCPR 2)
4. Freedom of opinion, thought, conscience, and religion (UDHR 18, 19; ICCPR 18, 19)
5. Freedom of assembly and association (UDHR 20; ICCPR 21, 21)
6. Privacy (UDHR 12; ICCPR 17)
7. Freedom from arbitrary arrest, detention, and exile; right to a fair trial; assumption of innocence before being proven guilty (UDHR 9, 10, 11;ICCPR 14; CERD 5)

Whose Human Rights

In order to draw a human rights lifecycle, it is also important to mention whose rights (as well as what rights are affected). Most of the following are taken from the BSR Human Rights Impact assessment report (we added

1 Dunstan Allison Hope, Lindsey Andersen, and Susan Morgan, "Human Rights Assessment, Global Internet Forum to Counter Terrorism," Business for Social Responsibility, July, 20, 2021, https://www.bsr.org/en/our-insights/report-view/human-rights-impact-assessment-global-internet-forum-to-counter-terrorism

"the accused" to the group of people whose rights might be affected).[2] We decided that economic and social groups might be impacted in the longer term and do not squarely fit the crisis protocol framework, so we might not include formal, societal, practical discrimination and hidden groups in the lifecycle.

1. **Victims of terrorism and violent extremism**

These rightsholders are the direct victims of terrorist activities that have an online angle. They are usually the ones that are the targeted group or the casualty.

2. **Victims of efforts to counter terrorism and violent extremism**

The victims of efforts to counter terrorism include groups that are adversely impacted by efforts to counter terrorism and violent extremism online. These groups are usually subject to overbroad or wrongful content removal or other actions. Another group is those who do the counterterrorism activities such as law enforcement, tech company security research departments, and those who undertake OSINT. Exposure to terrorist materials and undertaking research about these issues can affect the rights of these groups.

3. **Human rights defenders**

Human rights defenders include professional and citizen journalists, civil society organizations, nonviolent political activities, and members of vulnerable groups advocating for their rights.

4. **Women, girls, families, men, and boys**

According to the BSR report women, girls, and families as well as men, boys and the LGBTQI+ community can be disproportionately hampered by terrorist and extremist content, for example arising from problematic use of gender stereotypes in efforts to counter terrorism and violent extremism.[3]

5. **The accused**

The accused is the potential perpetrator that might be identified as the person behind spreading terrorist and violent extremist content.

The list of impacted rightsholders is not exhaustive, as it is difficult to identify all the rightsholders which vary across geographies and context. We will however try as much as possible to discuss which rights of these rightsholders could be hampered during the crisis protocol.

Scope

The scope of the work and online activities are limited to online content incidents, involving livestream/video/audio/image/text that records or depicts a recent, ongoing or imminent real-world terrorist and violent extremist attack. It is also important to note that this document is concerned with the "crises" that have an online angle and do not include stages of radicalizations or other aspects.

2 Hope, Andersen, & Morgan, "Human Rights Assessment."

3 Hope, Andersen, & Morgan, "Human Rights Assessment."

Categories of Attacks

Based on a combination of the Global Terrorism Database and other categorizations at Council of Foreign Relations,[4] we have included the following aspects for categorizing terrorist attacks with an online aspect:

1. Geographical scope

Whether the attack crossed international borders or if there were citizens of different countries affected by the attack.

2. Used online service providers and online materials

What platforms were used and what content was shared online.

3. Multiple platforms

As part of the attack whether attackers or supporters exploited multiple platforms to share terrorist content online.

4. Virality

Defined as achieving a large number of views in a short time period due to sharing.[5]
Virality is maximized to the extent that content viewed by one consumer is shared with others.[6]

5. Type of attack and weapon information

Whether the terrorist attacks included the use of weapons and what type was used.

6. Target type

Whether religious organizations, business institutions, government entities were the target.

7. Terrorist group name and kind

What kind of a terrorist group it was and whether it state sponsored or not.

8. Number of perpetrators

How many perpetrators and accomplices undertook the attack.

9. Number of casualties

The number of people who were injured or killed.

10. Claims of responsibility

Whether a group claimed responsibility for the attack.

4 Cyber Operations Tracker, Council on Foreign Relations, n.d., https://www.cfr.org/cyber-operations/.

5 Alhabash and McAlister see virality as a combination of user-generated activities performed on social networks. They take a behavioral approach to defining virality by focusing on viral reach (i.e., access to and sharing of content), affective evaluation (i.e., likes and dislikes), and message deliberation (i.e., comments and status updates). See Saleem Alhabash and Anna R. McAlister, "Redefining virality in less broad strokes: Predicting viral behavioral intentions from motivations and uses of Facebook and Twitter," New Media & Society 17, no. 8 (2015): 1317–1339.

6 Gerard Tellis et al., "What drives virality (sharing) of online digital content? The critical role of information, emotion, and brand prominence," Journal of Marketing 83, no. 4 (2019): 1–20.

11. Timing

If the attack happened in the past, is ongoing or imminent.

12. Threats

Terrorist threats that did not materialize or were prevented/not undertaken.

13. Intention for mass violence

Whether the terrorist has a clear intention to undertake mass violence.

14. Shared by (perpetrator, accomplice, sympathizer, bystander)

Who shared the materials and how were they shared.

15. Used crisis protocols

Which crisis protocol was activated or is about to be activated.

16. Prior conviction

Whether the perpetrator(s) had a prior conviction or arrest on related issues.

Categories	Halle	Glendale	Conflans	London Bridge	Vienna	Washington DC	Streatham	Reading	Christchurch
Geographical scope	Global	Global	Global	Local	Local	Local	Local	Local	Global
Used Online Service providers	Livestream on Twitch, Manifesto on Meguca	Live streamed, Snapchat	Twitter, footage	Videos and Pictures	Bystander videos and subsequent attacker manifesto release after the incident (video)	Parler, Twitter, Facebook, GAB	Bystander video of the police response	Bystander video of the aftermath shared	Initially, 8Chan with links to livestream on Facebook, and to manifesto on Mega, Solidfiles, Zippyshare, Mediafire
Multiple platforms	Yes	Yes	Yes	Unknown	No (manifesto was shared more widely)	Yes	Unknown	Yes (Instagram and Twitter)	Yes
Virality	5 livestream views; 2200 recorded views	No	Image posted on Twitter, liked by some	Unknown	Unknown	Yes	No	No	Yes
Type of attack	Armed Assault	Armed Assault	Armed Assault	Armed Assault	Armed Assault	Armed and Unarmed Assaults	Armed Assault	Armed Assault	Armed Assault
Weapon information	Handmade gun	Gun	30cm knife	Knife	Gun	Use of various weapons	Knife	Knife	Gun
Target type	Religious institution, Business	Business	Civilian, Educational institution	Civilians	Civilians	Government officials and buildings	Civilians	Civilians	Religious institution (mosques) and Muslim Civilians
Terrorist group name or kind	Anti-Semitic extremist	Incel	Radicalized individual (Jihadi inspired)	Jihadi Inspired Extremism	Alleged IS Supported attack	Anti Government/ Authority	Jihadi Inspired terrorism	ASL (AQ aligned group based in Libya)	White identity motivated violent extremism
Number of perpetrators	1	1	1	1	1	725 arrests	1	1	1
Number of casualties	1 killed	3 injured	1 killed	3 killed (including perpetrator)	5 killed, 23 injured	7 killed, many injured	3 injured	3 killed, 3 injured	51 killed, many injured
Claims of responsibility	N/A	N/A	No (however the perpetrator has featured in ISIL propaganda, was in contact with ISIL and his brother was a member of ISIL)	No	Yes (Amaq)	No (some individuals have accepted responsibility during their individual trials or plea agreements)	No (attacker had previously been convicted of offenses relating to disseminating AQ material)	As well as being a member of ASL, the attacker had ISIL material on his device	N/A
Timing	Past	Past	Past	Past	Past	Past	Past	Past	Past
Threats	No	No	No	No	No	No	No	No	No
Intention for mass violence	Yes (though did not fully materialize)	Yes	No	Yes	Yes	Unknown	Yes	Yes	Yes
Shared by	Perpetrator	Perpetrator	Perpetrator	Bystanders	Bystanders	Perpetrators, Accomplices, Sympathizers, Bystanders	Bystanders	Bystanders	Perpetrator
State responsibility	No	No	No	No	No	No	No	No	No
Used crisis protocols	GIFCT (CIP)	GIFCT (CIP)	EU	UK	New Zealand (not fully activated but the monitoring stage triggered)	None	UK	UK	No (was prior to (and reason for) development of all except UK protocols
Prior conviction	No	No	Unrelated charges	Yes	Yes	Yes for some perpetrators	Yes	Yes	No

Case Studies

We will briefly discuss the human rights implications of the following attacks at each stage since they resulted in monitoring the attack or activation of various Crisis Response Protocols:

- Halle, Germany, October 2019
- Christchurch, New Zealand, March 2019
- Glendale, AZ, USA, May 2020
- Conflans, France, October 2020
- Washington, D.C., USA, January 2021

As much as possible, the cases are studied by following the Crisis Protocol stages. However in some cases, some stages have been collapsed into one either because there was not much information about activities during that stage or the activities could not be analyzed based on each stage.

Halle

This incident was an armed attack on a synagogue. The perpetrator posted a manifesto (a platform called Meguca which is loosely affiliated with 4Chan) detailing the attack.[7] During the horizon period, there was no monitoring or surveillance. The perpetrator recorded himself from the beginning of the attack in his car, streaming on Twitch for 35 minutes, with approximately five people viewing it live and 2200 people viewing the recording of it. After somebody reported the "recorded" video, it was taken down.[8]

In its transparency report on the incident GIFCT stated the following: "On Wednesday, October 9, 2019, the GIFCT activated its new Content Incident Protocol (CIP) for the first time after the protocol's development following the terrorist attack in Christchurch, New Zealand the previous March. The CIP was declared following the tragic shooting in Halle, Germany and the perpetrator's attack video circulating on multiple digital platforms."[9]

Stages 1-4: Horizon, Identify and validate, Incident detection, and Information gathering and validation:

1. **Life, liberty, and security of person –** During the horizon stage, the act of live streaming had the potential to incite more violence. It was, however, not viral content. The rightsholders were actual and potential victims of terrorism and violent extremism as the video streaming could incite more violence. We are not aware of any other action that had an impact on other human rights and rights holders at this stage.
2. **Nondiscrimination and equality before the law –** No known violations.
3. **Access to effective remedy –** No known violations.
4. **Freedom of opinion, thought, conscience, and religion –** No known violations.
5. **Privacy –** No known violations.
6. **Freedom from arbitrary arrest, detention, and exile; right to a fair trial; assumption of innocence**

7 Daniel Koehler, "The Halle, Germany, Synagogue Attack and Evolution of the Far-Right Terror Threat," CTC Sentinel 12, no. 11 (2019), https://ctc.westpoint.edu/wp-content/uploads/2020/02/CTC-SENTINEL-112019.pdf.

8 See Twitch's Twitter thread on the 2019 Halle incident: https://twitter.com/Twitch/status/1182036266344271873.

9 Update to GIFCT Statement on Halle Shooting, GIFCT, October 17, 2019, https://gifct.org/2019/10/17/update-to-gifct-statement-on-halle-statement/.

before being proven guilty – No known violations.

Stage 5: Assess

1. **Life, liberty, and security of person –** No known violations.
2. **Nondiscrimination and equality before the law –** No known violations.
3. **Access to effective remedy –** No known violations.
4. **Freedom of opinion, thought, conscience, and religion –** No known violations.
5. **Privacy –** No known violations.
6. **Freedom from arbitrary arrest, detention, and exile; right to a fair trial; assumption of innocence before being proven guilty –** No known violations.

Stage 6: Activate and notify

During this stage a Twitch statement was released, but it is not clear what level of action GIFCT members were told should be taken.

1. **Life, liberty, and security of person –** At this stage, the live streaming and the video could have a high impact on life, liberty, and security of a person, had the content gone viral. But since it did not, the impact on this human right is unknown.
2. **Nondiscrimination and equality before the law –** No known violations.
3. **Access to effective remedy –** Unlikely to have been impacted at this stage; however, if the content was not preserved as a result of protocol operator guidance, it could affect access to effective remedy.
4. **Freedom of opinion, thought, conscience, and religion –** No known violations.
5. **Privacy –** No known violations.
6. **Freedom from arbitrary arrest, detention, and exile; right to a fair trial; assumption of innocence before being proven guilty –** No known violations.

Stage 7: Prepare and Act

1. **Life, liberty, and security of person –** This stage had a high impact on life, liberty, and security of person; despite the fact that the content was found at a very late stage, the actions by the protocol operators potentially stopped it from going viral.
2. **Nondiscrimination and equality before the law –** No known violations.
3. **Access to effective remedy –** This stage potentially had a high impact on access to effective remedy as it relates to preserving evidence.
4. **Freedom of opinion, thought, conscience, and religion –** No known violations.; at this stage the hash database might not have a high impact on freedom of opinion, thought, conscience, and religion, but at a later stage, if hashes are inaccurate it could have an impact on human rights in the future.
5. **Privacy –** No known violations.; however, gathering OSINT materials can lead to profiling.
6. **Freedom from arbitrary arrest, detention, and exile; right to a fair trial; assumption of innocence before being proven guilty –** No known violations.

Stage 8: Conclude

There might be limited human rights implications during this stage; however, the decisions that are taken based

on the lessons learned from the attack (like expanding the hash database) could have future human rights implications. For this report , we are not aware of cases of human rights impact at this stage.

Glendale (The Westgate Shooting)

An involuntary celibate (incel) sympathizer shot couples at a mall. There was an online component to this, as the GIFCT transparency report mentions: "On Wednesday, May 20, 2020, the GIFCT activated its Content Incident Protocol following the shooting in Glendale, AZ, adding hashes of visually distinct videos depicting the attacker's content during the shooting."

In Snapchat videos released by police, the perpetrator said he was going to be the shooter, along with another clip showing his gun where he says, "Let's get this done, guys." In this case, the perpetrator did not have any criminal background. The protocol was not activated during the incident and it seems to be an after the fact reaction to the incident.[10]

Stage 1: Horizon

The perpetrator streamed the video on Snapchat shortly before the incident. There was no monitoring and there was no reporting. It is not publicly known if the brother reported him or the content to the police.

Because the video was found after the incident, there were no known violations. of human rights implications. There might have been some greater human rights implications during the horizon that could potentially incite others, such as members of Incel groups or the actor's followers.

1. **Life, liberty, and security of person –** Nobody died because of streaming, but it had the potential to incite further violence.
2. **Nondiscrimination and equality before the law –** No known violations.
3. **Access to effective remedy –** No known violations.
4. **Freedom of opinion, thought, conscience, and religion –** No known violations.
5. **Privacy –** No known violations.
6. **Freedom from arbitrary arrest, detention, and exile; right to a fair trial; assumption of innocence before being proven guilty –** No known violations.

Stage 7: Prepare and Act

Since hash sharing occurs at this stage, it could have had some human rights implications on future events. In this case, the information sharing led to the expansion of hash databases that could have potential implications for human rights.

1. **Life, liberty, and security of person –** The ongoing information sharing and sharing hashes might have prevented the content from going viral.
2. **Nondiscrimination and equality before the law –** No known violations.
3. **Access to effective remedy –** No known violations.
4. **Freedom of opinion, thought, conscience, and religion –** No known violations.

10 "CIP declared following Glendale, AZ shooting," GIFCT, May 21, 2020, https://gifct.org/2020/05/21/cip-declared-following-glendale-az-shooting/.

5. **Privacy –** No known violations.
6. **Freedom from arbitrary arrest, detention, and exile; right to a fair trial; assumption of innocence before being proven guilty –** No known violations.

Stage 8: Conclude

Depending on what kind of information GIFCT and others shared with each other, this stage could have different human rights implications. There are no known violations. of impact on human rights.

Samuel Paty, France, Conflans

Samuel Paty was a teacher at a high school in Conflans. He allegedly showed cartoons of the prophet Muhammad to students. A parent of a student filed a criminal complaint with police. On Youtube and Facebook the parents claimed that Paty displayed an image of Muhammad, named Paty in the video and gave the school address. An Imam posted a video on a social media platform and called Paty a thug. In October 2020, the perpetrator saw the video made by the imam and decided to punish Paty. Minutes after the attack, the perpetrator posted a picture of Paty's severed head on Twitter. The picture was seen by many of Paty's students. The EU Crisis Protocol was activated.

Stages 1-4: Horizon, Identify and validate, Incident detection, and Information gathering and validation

Before the attack, in the horizon phase, it is unclear whether there was monitoring and surveillance by those in charge of CIP. This stage has high human rights implications, the right to privacy, assembly but also security, liberty, and freedom are just a few that can be hampered.

1. **Life, liberty, and security of person –** Monitoring and surveillance (if targeted and proportional) might have helped protect Paty's right to life, liberty, and security.
2. **Nondiscrimination and equality before the law –** No known violations.
3. **Access to effective remedy –** No known violations.
4. **Freedom of opinion, thought, conscience, and religion –** If monitoring took place it could lead to arrest or also hampering assembly rights (the right for Muslims to organize a protest online against the teachers or express their beliefs through online content).
5. **Privacy –** At these preactivation stages, the identification and validation must have been easier and more straightforward since the action was materialized and the picture was posted. Because of these factors, there might have been a low level of human rights implications.
6. **Freedom from arbitrary arrest, detention, and exile; right to a fair trial; assumption of innocence before being proven guilty –** No known violations.

Stage 7: Prepare and Act

It does not look like that the image was posted across different platforms, but removal of content at this stage (instead of making it private) might make future investigations harder. The rapid deletion by the platforms of the images of the attack have indeed hindered police investigations by depriving them of visual information on the terrorist and the place to which he headed after the attack for example. This important information could not be provided to the police. This could have hindered human rights by allowing the terrorist to carry out further attacks before his arrest. It is therefore important to take this into account when assessing the human rights impacts of crisis protocols. A suppression in the public sphere but the preservation of the files by the platforms

would be relevant in this sense.

This incident might have led to the creation of new hashes. Using OSINT by the CIP could lead to finding others in the friends circle which could hamper privacy, freedom of opinion and thought. It is unclear if the OSINT that were used during the crisis led to the arrest of others, but it is a human rights implication that should be considered.

1. **Life, liberty, and security of person –** Activation happened after Paty was murdered; no other threat to life, liberty, and security as a result of activation could be predicted in this case. Paty's picture was posted by the terrorist on Twitter and was seen by students before it could be taken down. While this is not the direct result of late activation of protocol, potentially Paty's students' right to security could be hampered.
2. **Nondiscrimination and equality before the law –** No known violations.
3. **Access to effective remedy –** Removal of materials made it difficult to investigate the attack and provide effective remedy.
4. **Freedom of opinion, thought, conscience, and religion –** No known violations.
5. **Privacy –** Using OSINT by CIP operators could lead to finding others in the friends circle which can hamper privacy, freedom of opinion and thought. In fact a person who liked the terrorist tweet was arrested.
6. **Freedom from arbitrary arrest, detention, and exile; right to a fair trial; assumption of innocence before being proven guilty –** It does not seem that anybody was arrested as a direct result of activation of the protocol. There were some arrests made by the French police.

Stage 8: Conclude

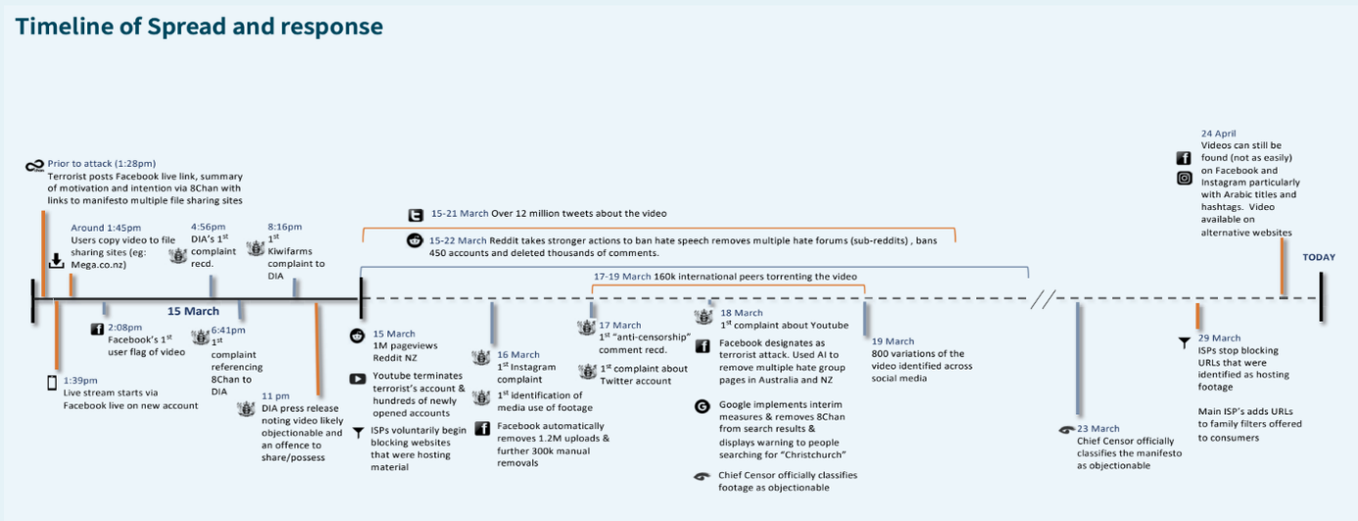
The conclude phase did not happen at GIFCT because the protocol was not activated, but GIFCT could have benefited from learning from this experience and debriefing the stakeholders which could contribute to life, liberty, security as well as privacy. Since monitoring continues, despite the lower human rights implications, there might be more human rights implications in the future. A year after the attack, a blogger talked to the father of the perpetrator who endorsed the acts of his son. Somebody who liked the decapitated image of Paty on Twitter was also arrested.

Following the attack on Samuel Paty, the offense of endangering the life of others by providing personal data was created by law no. 2021-1109 of August 24, 2021, reinforcing the respect of the principles of the Republic, article 36. A penalty of 3 years imprisonment and 45,000 euros fine was created (5 years and 75,000 euros if the victim is a minor or a representative of the public authority, a person in charge of a public service mission, or has an elective mandate). It is now illegal to disclose someone's name and home or work address while calling for online hate or violence against them. This provision aims at monitoring and condemning this type of publication while trying to preserve freedom of speech online. This is not a protocol but a measure that can help avoid dramatic situations such as the one in Conflans.

Christchurch Attack (hypothetical analysis)

No multi-party protocols existed at the time of the Christchurch terrorist attack (other than Facebook's own three step crisis protocol), so what follows here is a hypothetical analysis. In an extensive violent extremism transparency report, New Zealand Department of Internal Affairs (DIA), provided details and a timeline about the

attack and its online angle. The chart below is taken from the transparency report that the DIA published in April 2022.[11]



Stages 1-4: Horizon, Identify and validate, Incident detection, and Information gathering and validation

The perpetrator posted an anonymous message to an online discussion board called 8chan and revealed his intentions to undertake an attack and livestream it. There was a link to his Facebook page that was repeatedly shared. He also sent messages and emails to family and the New Zealand Prime Minister's office. He started live streaming as he went towards Masjid an-Nur.

1. **Life, liberty, and security of person –** This stage clearly had an impact on life, security and liberty of the Muslims who were murdered; the live streaming at this stage could lead to other incidents as well.
2. **Nondiscrimination and equality before the law –** Disproportionate focus on Islamic Extremism resulted in discrimination against Muslims.
3. **Access to effective remedy –** No known violations.
4. **Freedom of opinion, thought, conscience, and religion –** No known violations.
5. **Privacy –** No known violations.
6. **Freedom from arbitrary arrest, detention, and exile; right to a fair trial; assumption of innocence before being proven guilty –** No known violations.

Stage 5: Assess

1. **Life, liberty, and security of person –** This was a clear cut terrorist act, and the assessment at this stage would have had high impact on life, liberty, and security of person, since it had the potential to go viral (and it did), and had a clear terrorist violent intention.
2. **Nondiscrimination and equality before the law –** Biases observed in the development and implementation of counter terrorist policy must be guarded against, and assessments and decisions

11 New Zealand Department of Internal Affairs, "2021 Digital Violent Extremism Transparency Report," April, 2022. https://www.dia.govt.nz/diawebsite.nsf/Files/Countering-violent-extremism-online/$file/DVE-Transparency-Report-2021-a.pdf.

made must be able to show that biases were considered and addressed.

3. **Access to effective remedy –** No known violations.
4. **Freedom of opinion, thought, conscience, and religion –** Some users who shared the video in good faith to spread awareness about the incident felt that assessing their post as spreading extremist content was against their freedom of expression; freedom of religion could be impacted if assessment of materials have a lower threshold (as it instills fear in people).
5. **Privacy –** No known violations.
6. **Freedom from arbitrary arrest, detention, and exile; right to a fair trial; assumption of innocence before being proven guilty –** No known violations.

Stage 6: Activate and notify

1. **Life, liberty, and security of person –** As the perpetrator made his way to another Islamic Center, activation of protocol and informing civil society groups about it might have helped by warning the Muslim community about the online material and the incident, protecting security, life and liberty at this stage.
2. **Nondiscrimination and equality before the law –** No known violations.
3. **Access to effective remedy –** No known violations.
4. **Freedom of opinion, thought, conscience, and religion –** No known violations.
5. **Privacy –** No known violations.
6. **Freedom from arbitrary arrest, detention, and exile; right to a fair trial; assumption of innocence before being proven guilty –** No known violations.

Stage 7: Prepare and Act

1. **Life, liberty, and security of person –** In the future, sharing hashes at this stage might reduce the virality and cross platform movement of the terrorist content.
2. **Nondiscrimination and equality before the law –** No known violations.
3. **Access to effective remedy –** No known violations., but if there were mistakes in taking down the content without preserving it, it could hamper access to effective remedy (since the terrorist content went viral and shared on multiple platforms, hypothetically not all the content could be taken down and some could be preserved).
4. **Freedom of opinion, thought, conscience, and religion –** No known violations.
5. **Privacy –** No known violations.
6. **Freedom from arbitrary arrest, detention, and exile; right to a fair trial; assumption of innocence before being proven guilty –** No known violations.

Stage 8: Conclude

The online angle of Christchurch terrorist attack led Prime Minister of New Zealand and President of France to launch the Christchurch Call to Act in May 2019. The Call led to strengthening and expansion of the GIFCT and multi-stakeholder efforts to develop and implement protocols (including the Christchurch Call Crisis Response Protocol, the GIFCT Content Incident Protocol and Incident Response Framework, and the EU Crisis Protocol), enabling a rapid, coordinated, and effective response to the dissemination of terrorist or violent extremist content following a real-world attack.

Some tech companies after the attack took measures such as removing 8Chan from search results or displaying warnings about the Christchurch attack. Tech companies also made safety improvements to livestream

services.[12]

U.S. Capitol Attack

Stage 1: Horizon

A large crowd gathered outside the U.S. Capitol during a Joint Session of Congress which began at approximately 1:00 PM Eastern Time (ET). The FBI had a command post operation based out of Headquarters in Washington, D.C. in support of preventing violence and criminal activity in the National Capital region. Law enforcement agencies could see social media posts about the event, and mainstream news coverage on the major networks and local channels.

1. **Life, liberty, and security of person –** People had the choice to gather outside the U.S. Capitol during the Joint Session of Congress. U.S. Capitol Police (USCP) were present as it is their responsibility to keep members of Congress and Senators safe, as well as the public visiting U.S. Capitol grounds. In addition, USCP is responsible for protection of government property. While law enforcement agencies used tools to monitor social media platforms right before the incident, no action was taken with regards to content moderation in coordination with law enforcement agencies. At this stage, life, liberty, and security of persons such as the government representatives and the public could have been endangered.
2. **Nondiscrimination and equality before the law –** Any person of any race, creed, gender, etc., could gather outside the U.S. Capitol and each person is promised to be treated equally by Capitol Police regardless of race, creed, gender, etc. On social media, as they are private entities, the obligation of equal treatment does not legally exist; however, some tech-corporations have their own human rights policies. There are no known violations. of discrimination based on creed, gender etc. that happened during this stage.
3. **Access to effective remedy –** N/A at this stage of the event.
4. Freedom of opinion, thought, conscience, and religion – There are conflicting reports that some law enforcement agencies monitored online activities. This could potentially have had an effect on freedom of opinion, thought, conscience, and religion. However, there are reports that the law enforcement missed the threats on social media platforms and was not prepared and therefore all were free to express those opinions through social media.[13]
5. **Privacy –** People had the choice whether to gather outside the U.S. Capitol during the Joint Session of Congress on public grounds, and whether to publicly post videos or images of themselves inside the Capitol. If law enforcement undertook monitoring of the online activities, despite them being public, it could affect privacy of people in the later stages of the protocol.
6. **Freedom from arbitrary arrest, detention, and exile; right to a fair trial; assumption of innocence before being proven guilty –** N/A at this stage of the event.

Stage 2: Identify and Validate

At approximately 2:00 PM EST, individuals in the crowd forced their way through the barricades and the crowd advanced to the U.S. Capitol while the Joint Session was still underway. Law enforcement agencies evaluated

12 Ben Collins, "Facebook to restrict livestream feature after Christchurch attack," NCBCNews, May 14, 2019, https://www.nbcnews.com/tech/tech-news/facebook-restrict-livestream-feature-after-christchurch-attack-n1005741.

13 Sam Levin, "US Capitol attack: is the government's expanded online surveillance effective?," January 7, 2022, https://www.theguardian.com/us-news/2022/jan/07/us-capitol-attack-government-online-surveillance.

public social media posts for threats or indications of violent activity or violations of federal criminal law.

1. **Life, liberty, and security of person –** U.S. Capitol Police were unable to fulfill their responsibility to keep members of Congress and Senators safe, as well as the public visiting U.S. Capitol grounds. This right was affected by the number of officers, their communications with one another, the number and quality of barricades, and the number of people gathered.
2. **Nondiscrimination and equality before the law –** Any person of any race, creed, gender, etc., could force his/her way through the barricades and advance with the crowd at this stage of the event. Law enforcement agencies' evaluation of public social media posts did not consider the poster's race, creed, gender, etc., and each post was evaluated as a stand-alone post, and in the context of the user's other posts.
3. **Access to effective remedy –** N/A at this stage of the event if rights were not harmed.
4. Freedom of opinion, thought, conscience, and religion – Even as people forced their way through the barricades and advanced with the crowd, the people choose to speak or not speak. All were free to express those opinions on the grounds of the U.S. Capitol. Opposite viewpoints, however, expressed within earshot of each other, could have implications, if that expression led to a physical altercation. All were free to express those opinions through social media.
5. **Privacy –** Once people crossed into the restricted space of the U.S. Capitol, they had a diminished expectation of privacy. The U.S. Capitol is a restricted building, and the Joint Session was closed to the public.
6. **Freedom from arbitrary arrest, detention, and exile; right to a fair trial; assumption of innocence before being proven guilty –** N/A to the online angle at this stage.

Stage 3: Incident Detection

At approximately 2:20 PM EST, the order was given for evacuation of the chambers by members of Congress and the Senate and the session was suspended until approximately 8:00 PM EST. The incident was detected by several law enforcement agencies through social media exploitation.

Life, liberty, and security of person – The USCP, members of Congress and Senators, and the trespassing public all still retained these rights, however it was difficult for USCP to keep everyone safe in the ensuing mayhem. There were difficulties in identifying violent extremist content which could potentially impact life, liberty, and security of a person and distinguish if from other opinions.

1. **Nondiscrimination and equality before the law –** This would have been difficult for USCP to discern during the chaos. It would also be difficult for law enforcement if they intended to work with tech-platforms to moderate content on social media platforms as well.
2. **Access to effective remedy –** Rights might not have been harmed just yet, as the incident is being detected and recognized as an incident.
3. **Freedom of opinion, thought, conscience, and religion –** The detection stage could create a lot of problems for freedom of opinion, thought, and conscience as the content being shared on social media platforms was not clear-cut terrorist content material; detecting and isolating terrorist content was very difficult.
4. **Privacy –** Once people crossed into the restricted space of the U.S. Capitol, they had a diminished expectation of privacy. The U.S. Capitol is a restricted building, and the Joint Session was closed to the public. Areas inside the U.S. Capitol might have restrictions posted prohibiting photography and audio

or video recording. Violators of that posted policy who were observed by the police could expose themselves to confiscation of their devices.

5. **Freedom from arbitrary arrest, detention, and exile; right to a fair trial; assumption of innocence before being proven guilty –** N/A to the online angle at this stage.

Stage 5: Assess

This incident appeared and has been shown for some people (not all) to have a domestic violent extremist motivation. Abundant online content was produced and posted and streamed by perpetrators and bystanders alike.

1. **Life, liberty, and security of person –** Once members of Congress and Senators were safe, USCP would have turned to the public inside the chamber and worked to clear spaces one room/hallway at a time. USCP officers would have initially begun with verbal commands, and if people did not comply, they may have been detained or arrested. Some would have likely complied or tried to comply and retreat outside the U.S. Capitol grounds.
2. **Nondiscrimination and equality before the law –** This would still have been difficult for USCP to discern during the assessment phase of the incident, likely from after the recognition of an incident (after 2:20 PM) up until the session was reconvened about 8 PM. Who had a right to be inside and who did not? Who should be evacuated to safety, and how could officers tell the difference? Would they use visual cues such as clothing and dress, weapons in hand, etc., to determine who was trespassing? Could they determine who was trying to escape? This also applies to social media platforms and the same challenges could have been faced when working with that sort of content.
3. **Access to effective remedy –** Once a cell phone or other device has been confiscated, either incident to arrest, or for photographing in a prohibited area, it would require paperwork to recover that device. This process might seem arduous from a human rights perspective.
4. **Freedom of opinion, thought, conscience, and religion –** Because of the problems with discerning protest content and violent extremist content, at this stage freedom of opinion, thought, and conscience could have been impacted.
5. **Privacy –** Once people crossed into the restricted space of the U.S. Capitol, they and the content they posted publicly on social media platforms had a diminished expectation of privacy. The U.S. Capitol is a restricted building, and the Joint Session was closed to the public.
6. **Freedom from arbitrary arrest, detention, and exile; right to a fair trial; assumption of innocence before being proven guilty –** If arrested, a person's cell phone or other technology device would likely be confiscated and inventoried. A search warrant would have to be obtained for police to search the device, unless exigent circumstances warranted otherwise. Wrongful assessment of the online materials as violent extremism and terrorism could lead to arbitrary arrest.

Stage 6: Activate and Notify

A report of shots fired on the floor of the U.S. Capitol was sent out across government agencies at approximately 3:00 PM EST. The Mayor of Washington, D.C. imposed a 6:00 PM EST curfew for the city. Since no protocol was activated, we can only speculate what could happen.

1. **Life, liberty, and security of person –** At this stage, life, liberty, and security of person could be impacted if violent, extremist content was being shared consistently; however, since a curfew was in place it might have had lower impact.

2. **Nondiscrimination and equality before the law –** No known violations.
3. **Access to effective remedy –** Since this was the notification stage, access to effective remedy was unlikely to be highly impacted.
4. **Freedom of opinion, thought, conscience, and religion –** During activation stage, increased monitoring can take place which affects freedom of opinion and thought.
5. **Privacy –** Might not be impacted.
6. **Freedom from arbitrary arrest, detention, and exile; right to a fair trial; assumption of innocence before being proven guilty –** might not be impacted.

Stage 7: Prepare and Act

The Washington, D.C. National Guard was activated at approximately 4:00 PM EST. Bomb technicians were deployed for the device near the DNC. The same evening, the FBI opened an investigation into the civil disorder and riot allegations associated with the events at the Capitol.

1. **Life, liberty, and security of person –** By that evening, the FBI had opened an official investigation into the events that had occurred in the U.S. Capitol. This opened up many more resources for law enforcement to locate and identify those who were inside restricted space at the U.S. Capitol. Law enforcement agencies used a variety of lawful investigative techniques during the initial 24–48 hours of investigation. Many techniques were intrusive and involved collection of personal cell phone location information, subscriber data, and even content of communications in the restricted space.
2. **Nondiscrimination and equality before the law –** Law enforcement agencies opened individual investigations on each person who may have broken laws, including Entering and Remaining in a Restricted Building; Disorderly and Disruptive Conduct in a Restricted Building; Violent Entry and Disorderly Conduct in a Capitol Building; Parading, Demonstrating, or Picketing in a Capitol Building. Each person was treated individually before the law. Law enforcement also worked with private threat intelligence firms and activists to crowdsource pictures and footage of those who were involved in some way with the protest (use of OSINT at this stage was very extensive).
3. **Access to effective remedy –** Each person arrested and charged could choose to have his/her day in court, or agree to plead to charges levied against him/her. See "Capitol Breach Cases" posted through the United States Department of Justice, https://www.justice.gov/usao-dc/capitol-breach-cases. However, because of the use of OSINT and social media profiling that could increase error in identification, access to effective remedies could be highly hampered.
4. **Freedom of opinion, thought, conscience, and religion –** The public and those who agreed with the protest could be affected by take-downs of online content.
5. **Privacy –** This incident was very public. People who are formally charged are named in publicly available documents posted on the Internet. Privacy was diminished when formal charges were filed. A lot of profiling took place using OSINT when law enforcement cooperated with OSINT providers. Despite the fact that the footage and information were public, the activities that took place during this stage could hamper privacy of social media users.
6. **Freedom from arbitrary arrest, detention, and exile; right to a fair trial; assumption of innocence before being proven guilty –** Depending on which actions are taken, if guidance about the violent, and terrorist nature of the online content is inaccurate, it might lead to wrongful arrest.

Stage 8: Conclude

Lists of arrested individuals were circulated for database checks. Numerous lists were shared between

government agencies and departments in the immediate aftermath of the unrest at the Capitol. Although the crisis concluded, this remains an active investigation. Not all people have been identified and located who were in the Capitol that day.

Human Rights Analysis for Each Stage of the CRP

In this section, we provide the human rights matrix for each stage of the crisis response protocol. In order to have a more accurate and useful matrix, we have listed some qualitative indicators that might be present at each stage of the protocol and impact certain human rights.

Qualitative indicators are actions or elements that impact human rights when tackling a crisis online. We have identified these indicators through attending the table top exercises arranged by GIFCT, undertaking the case studies, and through discussions with different stakeholders. Some of these qualitative indicators loosely fit the definition that the United Nations has provided: "a human rights indicator is defined as specific information on the state or condition of an object, event, activity or outcome that can be related to human rights norms and standards; that addresses and reflects human rights principles and concerns; and that can be used to assess and monitor the promotion and implementation of human rights."[14]

The qualitative Indicators are as Follows:

1. Assessment of violent extremist or terrorist content
2. Virality
3. Cross platform
4. Broadening GIFCT's scope (for example not including hate speech since it broadens scope)
5. Diversity of stakeholders' consultation
6. Monitoring
7. Use of OSINT that can lead to profiling
8. Probability of false positive
9. Verification of information (how it's being verified, who gave it etc)
10. Accuracy in identifying perpetrator-only content
11. Criteria to assess significance of online presence
12. Expansion of hash database
13. Accuracy and completeness of guidance given to the GIFCT members
14. Accuracy and completeness of information sharing (companies act based on their own policy)
15. Take down of content and other actions taken
16. Sharing hashes with a stakeholders
17. Mitigation plan with a human rights analysis for future events

Assessment of violent or extremist content

- **What is it?** Activities that GIFCT or protocol operators undertake in order to assess whether the incident involves violent extremist or terrorist content.

14 United Nations, Office of the High Commissioner, "Human Rights Indicators: A Guide to to Measurement and Implementation," December 5, 2012, https://www.ohchr.org/sites/default/files/Documents/Publications/Human_rights_indicators_en.pdf.

- **Whose rights?** Victims of terrorism, victims of counter terrorism activities, the accused, human rights defenders.
- **Which rights?** Life, liberty, and security of person, freedom of opinion, thought, conscience, and religion, freedom of assembly and association, freedom from arbitrary arrest, detention and exile.

Virality

- **What is it?** Virality is achieving a large number of views in a short time period due to sharing.
- **Whose rights?** Victims of terrorism, victims of counter terrorism efforts, vulnerable groups.
- **Which rights?** A viral terrorist content could have an impact on the right to life, liberty, and security of person and privacy.

Cross platform

- **What is it?** Cross platform terrorist content is a piece of content that has been shared across multiple platforms and not limited to just one platform. Cross platform content does not necessarily go viral, but it can have a potential impact on human rights. This is especially the case when all the platforms are members of GIFCT and as a result might take similar actions that impact content across different platforms.
- **Whose rights?** Victims of terrorism, victims of counter terrorism efforts, vulnerable groups, the accused.
- **Which rights?** Nondiscrimination, freedom of opinion, freedom of assembly and association, access to effective remedy.

Broadening GIFCT's scope

- **What is it?** Actions that a protocol operator or GIFCT take that affect the scope of the protocol and broadens it.
- **Whose rights?** Victims of counter terrorism efforts, users of platforms, vulnerable groups.
- **Which rights?** Life, liberty, and security of person, freedom of assembly and association, freedom of opinion, expression.

Diversity of stakeholders' consultation

- **What is it?** At various stages of crisis management, protocol operators talk to third parties from civil society or law enforcement. The third parties might be involved in notifying the protocol operators about a potential incident. Whether the operators talk to a diverse set of stakeholders or not can impact human rights.
- **Whose rights?** Minorities and vulnerable groups, victims of terrorism, victims of counter-terrorism efforts.
- **Which rights?** Life, liberty, security of person, access to effective remedy, freedom of opinions, right to participation.

Monitoring

- **What is it?** Monitoring data sources, searching for and analyzing information to provide situational awareness, or inform response options.
- **Whose rights?** Victims, the accused, minorities/vulnerable groups.
- **Which rights?** Life, security of person and liberty, arbitrary arrest and the right to fair trial, privacy.

Use of OSINT that can lead to profiling

- **What is it?** Open Source Intelligence techniques use the publicly available information on the Internet in

order to identify a violent situation, the attacker, or pattern of attack.

- **Whose rights?** The accused, human rights defenders, minorities/vulnerable groups.
- **Which rights?** Freedom of opinion, thought, conscience, and religion, freedom from arbitrary arrest, detention, and exile, privacy.

Probability of false positive

- **What is it?** During some of the crisis protocol stages, it is more probable that content or conduct is identified falsely as terrorist and violent extremist.
- **Whose rights?** The accused, the victims of counter terrorism activities, human rights defenders, minorities, vulnerable groups
- **Which rights?** Freedom of opinion, thought, conscience, and religion, Freedom of assembly and association.

Verification of information

- **What is it?** The protocol operator receives information from third parties about a possible terrorist, violent extremist event with an online angle. The operator has to verify that information, using specific verification tools.
- **Whose rights?** Victims of efforts to counter terrorism, human rights defenders, vulnerable/minority groups.
- **Which rights?** Freedom of opinion, thought, conscience, and religion, freedom of assembly and association, freedom from arbitrary arrest, detention, and exile, right to privacy.

Accuracy in identifying perpetrator-only content

- **What is it?** The protocol operator has to identify perpetrator-only content because the implications for bystanders who are platform users could be grave. For example it could lead to blocking their accounts for some time.
- **Whose rights?** Victims of efforts to counter terrorism, the accused, users of platforms and bystanders, human rights defenders.
- **Which rights?** Freedom of opinion, thought, conscience, and religion, Freedom of assembly and association.

Criteria to assess significance of online presence

- **What is it?** The criteria that the protocol operator uses to assess the significance of online presence. This could also be a part of assessing threshold.
- **Whose rights?** Victims of terrorism, victims of efforts to counter terrorism.
- **Which rights?** Right to life, security of person and liberty.

Expansion of hash database

- **What is it?** A database of distinct hashed images of terrorist and violent extremist content. The hash database could potentially lead to removal of content that is not terrorist and violent. Moreover, because the companies share the hashes with each other, if there is a mistake, it happens across platforms.
- **Whose rights?** Victims of terrorism, victims of counter terrorism efforts, human rights defenders, vulnerable communities.
- **Which rights?** Access to effective remedy, freedom of opinion, thought, conscience, and religion, freedom of assembly and association.

Accuracy and completeness of guidance given to the GIFCT members

- **What is it?** The protocol operator (in this case GIFCT) might give guidance to member companies about what sort of action should be taken.
- **Whose rights?** Victims, the accused, victims of efforts to counter terrorism, minorities and vulnerable groups, human rights defenders.
- **Which rights?** Life, liberty, and security of person, nondiscrimination and equality before the law, access to effective remedy, freedom of opinion, thought, conscience, and religion, freedom of assembly and association, privacy, freedom from arbitrary arrest, detention, and exile, right to a fair trial, assumption of innocence before being proven guilty.

Accuracy and completeness of information sharing

- **What is it?** The protocol operator shares information with member companies about the incident, especially the pattern and its characteristics on the Internet.
- **Whose rights?** Victims, the accused, victims of efforts to counter terrorism, minorities and vulnerable groups, human rights defenders.
- **Which rights?** Life, liberty, and security of person, nondiscrimination and equality before the law, access to effective remedy, freedom of opinion, thought, conscience, and religion, freedom of assembly and association, privacy, freedom from arbitrary arrest, detention, and exile, right to a fair trial, assumption of innocence before being proven guilty.

Take down of content and other actions taken

- **What is it?** The operator can recommend take-down or other actions such as de-amplification. Companies might have their own internal mechanisms to mitigate harm, however smaller tech-corporations might not have the resources and act based on the operator's recommendation, so it has potential implications for human rights.
- **Whose rights?** Victims, the accused, victims of efforts to counter terrorism, minorities and vulnerable groups, human rights defenders
- **Which rights?** Life, liberty, and security of person, nondiscrimination and equality before the law, access to effective remedy, freedom of opinion, thought, conscience, and religion, freedom of assembly and association, privacy, freedom from arbitrary arrest, detention, and exile, right to a fair trial, assumption of innocence before being proven guilty, right to participation.

Sharing hashes with a stakeholder

- **What is it?** Sharing hashes with a stakeholder (for example GIFCT sharing hashes with law enforcement) does not happen, but if in the future this becomes a practice it will have human rights implications.
- **Whose rights?** Victims, the accused, victims of efforts to counter terrorism, minorities and vulnerable groups, human rights defenders.
- **Which rights?** Life, liberty, and security of person, nondiscrimination and equality before the law, access to effective remedy, freedom of opinion, thought, conscience, and religion, freedom of assembly and association, privacy, freedom from arbitrary arrest, detention, and exile, right to a fair trial, assumption of innocence before being proven guilty.

Mitigation plan with a human rights analysis for future events'

- **What is it?** Mitigation plan that considers a human rights analysis for future events learns from the past

human rights violations during a just concluded crisis and tries not to make the same mistakes in future crises.

- **Whose rights?** Victims of terrorism and violent extremism, Victims of efforts to counter terrorism and violent extremism, Human rights defenders, minorities and vulnerable groups and the accused
- **Which rights?** Life, liberty, and security of person, Nondiscrimination and equality before the law, Access to effective remedy, Freedom of opinion, thought, conscience, and religion, Freedom of assembly and association, Privacy, Freedom from arbitrary arrest, detention, and exile; right to a fair trial; assumption of innocence before being proven guilty

1. Horizon: when the attack is about to happen

During the Horizon stage, when (for example) the terrorist streams right before undertaking the attack, tech-corporations and GIFCT might undertake situational awareness, which includes monitoring various platforms.

Whose rights and which rights?

Monitoring and surveillance have human rights impact on victims of terrorism and violent extremism (threat to life, security and liberty) and the streaming might incite more violence at that moment. During horizon, the virality of the content and if multiple platforms are used to stream and disseminate the information are two indicators that can affect human rights of the victims of terrorism and the victims of counterterrorist activities. If the protocol operators at this stage undertake research that is not within their mandate or do not consult with a diverse set of stakeholder groups, it might lead to broadening the scope of the protocol. Broadening the scope of the protocol can impact the rights of human rights defenders and victims of terrorist and violent extremist acts while the victims of counter terrorist and violent extremist activities might be violated.

Qualitative indicators:

- Monitoring
- Virality
- Cross platforms
- Broadening GIFCT's scope
- Diversity of stakeholders consultation

2. Identify and validate and Incident detection (through internal monitoring, or a tip received from a partner organization, or media reporting).Information gathering and validation/part of pre-activation (seeking information to understand what has happened or is happening and ensure that understanding is valid, i.e. corresponds to reality).

This step is when a tip is received by third parties and they identify and validate whether there is livestreaming/violent content. Monitoring the incident takes place at this stage.

Whose rights and which rights?

Victims of terrorism and violent extremism's right to life, liberty, and security of person might be hampered if the attack is not validated correctly. The rights to freedom of opinion, freedom of assembly and association, privacy and freedom from arbitrary arrest and detention of victims of efforts to counter terrorism and violent extremism might be violated if there is a false positive.

Verification of information received from the third party, where the third party verification comes from, is it a

trusted source and the methods used to verify the trustworthiness of the source are all indicators of human rights impact at this stage. These indicators can increase the likeliness of inaccurate validation and violate the rights of victims of terrorist activities, victims of counterterrorism activities, the accused and others.

Use of OSINT might affect the rights of the accused and minorities in case of profiling. Profiling can have human rights impacts such as unfair arrest and the right to fair trial.

Qualitative indicators:

- Monitoring
- Use of OSINT
- Probability of false positive
- Verification of information (how it's being verified, who gave the information etc)

3. Assess

At this stage efforts are made to determine whether the attempt is mass violence, has a significant online presence and is it by the terrorist or accomplice or by a bystander.

Whose rights and which rights?

Victims of terrorism and violent extremism right to life, liberty, and security of person might be violated if assessment of mass violence and significance of online presence is not accurate. Victims of efforts to counter terrorism and human rights defenders right to freedom opinion, assembly and association as well as privacy might be violated if mass violence is detected incorrectly.
Victims of efforts to counter terrorism right and the accused might be violated if bystander materials are identified as perpetrator's materials (because it might be taken down, which will remove evidence). The users of a platform right to assembly and opinion might be violated if bystander content is flagged (because it can lead to suspension and blocking of their account). Victims of terrorism and violent extremism right to life and liberty might be threatened by the bystander material.

Qualitative indicators

- Accuracy in identifying perpetrator-only content
- Criteria to assess significance of online presence
- Assessment of violent or extremist content
- Probability of false positive

4. Activate and notify

This stage is critical as the protocol operator activates the protocol, (and if the operator is GIFCT) informs the members (tech-corporations) about the incident and informs them of the level of action needed.

Whose rights and which rights?

During the activation stage, the protocol operators will reach out to their members and other stakeholders with information about the attack. As the activate and notify stage provides guidance to various members about the level of action that is needed, it has a potential high impact on human rights.

The accuracy of information they have received, the use of OSINT and accuracy of the assessment will have an impact on human rights. Victims of terrorist activities' right to life, liberty, and security of person might be violated

in case of wrong assessment of the event and lack of action. For example, if live streaming is not interrupted, it might potentially impact the right to life and liberty. If a wrong assessment was made and the material was not violent extremist, it could lead to broadening the scope of the protocol. This could have an impact on human rights defenders that might have their content taken down or users that might have their accounts blocked (right to opinion and freedom of expression).

Qualitative Indicators:

- Accuracy in identifying perpetrator-only content
- Criteria to assess significance of online presence
- Assessment of violent or extremist content
- Probability of false positive

5. Prepare and Act (information sharing stage):

In prepare and act stage the protocol operators look at Open Source Intelligence materials, share hashes with their members (usually tech corporations), share awareness about where the violent extremist and terrorist content is, share the outcome of assessment, and engage in ongoing strategic communications.

Whose rights and which rights?

Victims of terrorism and violent extremism right to life, liberty, and security of person and right to effective remedy might be violated in case of transmitting wrong information that could lead to the deterioration of the situation. For example, if content is taken down, the terrorists might become more violent and kill more people. Or at this stage if content is taken down without preservation then the right to effective remedy might be violated.

Victims of efforts to counter terrorism and violent extremism right to freedom of opinion, thought, conscience, and religion, as well as freedom of assembly and association, might be violated due to take-down, deamplification and suspension of accounts. The right to privacy might also be affected due to use of OSINT (since it can lead to profiling) and hash-sharing. Human rights defenders' right to freedom of assembly and association as well as freedom of opinion might also be violated if actions such as hash-sharing lead to take-down and deamplification.

Qualitative Indicators:

- Expansion of hash database
- Accuracy and completeness of information sharing
- Take down of content
- Actions taken other than content take-down
- Sharing hashes with third parties

Conclude

Assessment against threshold, monitoring goes down, and producing summaries of what's gone on.

Whose rights and which rights?

The kinds of actions that are taken during this phase can have future human rights implications of operating a crisis protocol. In conclude the operators should pay special attention to how their actions impacted human rights at each stage of the crisis protocol and come up with a mitigation plan for future incidents.

Qualitative indicators may include:

- Diversity of stakeholder consultation
- Mitigation plan with a human rights analysis for future events

Human Rights Matrix

This human rights matrix maps out and evaluates the impact of the crisis protocol at each stage on human rights. This method needs to be polished and improved upon but it can potentially illustrate how and why each stage can impact human rights.

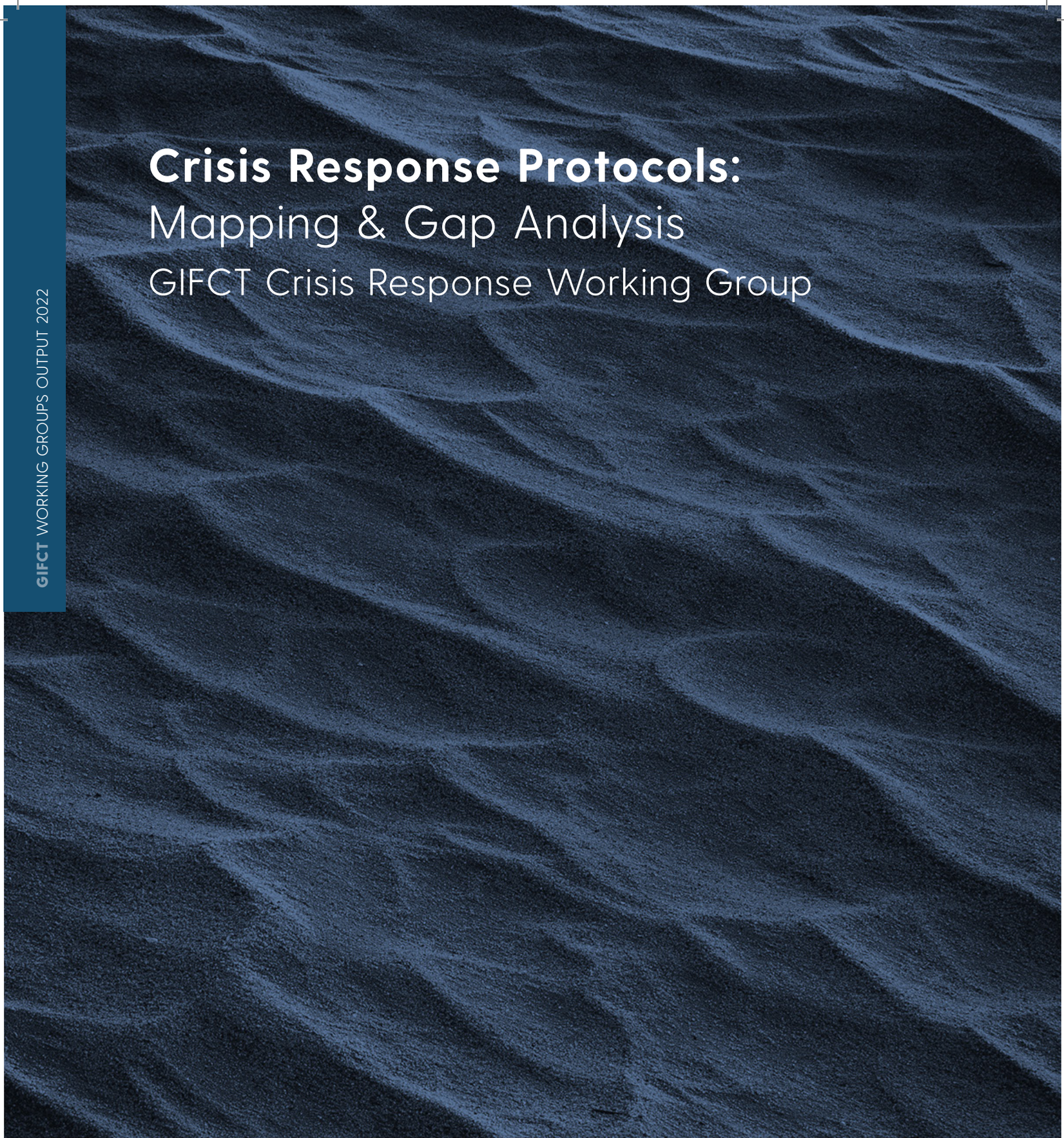
Stages of crisis and protocol	Whose and which rights?	Which rights?	Qualitative indicators
Horizon	During the Horizon stage, when the terrorist start the streaming right before undertaking the attack, tech-corporations and GIFCT might undertake situational awareness which includes monitoring various platforms. Monitoring and surveillance has human rights impact for victims of terrorism and violent extremism (threat to life, security of person and liberty) because the streaming might incite more violence at that moment. The victims of efforts to counter terrorism might also have their right to privacy and freedom of opinion violated at the horizon stage, in case there is a false positive. Vulnerable groups (women, girls, men, families) could have their right to liberty, and security of person violated. Human rights defenders: human rights defenders could be violated if they are reporting on police brutality or other events which could potentially be falsely flagged as terrorist content or violent extremist content Accused rights, in case of false positives, at this stage because situational awareness is happening and the protocol scope might be expanded if the attack is out of scope, it is possible that the accused right to freedom of opinion be violated.	Life, liberty, and security of person Access to effective remedy Freedom of opinion, thought, conscience, and religion Freedom of assembly and association Privacy Freedom from arbitrary arrest, detention, and exile;	- Situational awareness - Virality - Multiple platforms - Broadening GIFCT's scope - Monitoring and surveillance
Identify and validate	Victims of terrorism and violent extremism right to life, liberty, and security might be violated if the attack not validated correctly. Victims of efforts to counter terrorism and violent extremism rights to freedom of opinion, freedom of assembly and association, privacy and freedom from arbitrary arrest and detention might be violated if there is false positive. Use of OSINT might affect the rights of the accused and minorities in case of profiling, right to privacy	Life, liberty, and security of person Nondiscrimination and equality before the law Access to effective remedy Freedom of opinion, thought, conscience, and religion Freedom of assembly and association Privacy Freedom from arbitrary arrest, detention, and exile; right to a fair trial; innocence,before being proven guilty	- Monitoring - Probability of false positive - Verification of information (how it's being verified, who gave it etc) - Diversity of stakeholders' consultation - Use of OSINT

Assess: whether the attempt is mass violence, has a signif-icant online presence, is it by the terrorist or accomplice or by bystander.	Victims of terrorism and violent extremism right to life, liberty, and security might be violated if assessment of mass violence and significance of online presence is not accurate. Victims of efforts to counter terrorism and human rights defenders right to freedom opinion, assembly and association as well as privacy might be violated if mass violence is detected incorrectly. Victims of efforts to counter terrorism right and the accused might be violated if bystander materials are identified as perpetrator's materials. (because it might later on be taken down, which will remove evidence) The users of a platform right to assembly and opinion might be violated if bystander content is flagged (because it can lead to suspension and blocking of their account) Victims of terrorism and violent extremism right to life and liberty might be threatened by the bystander material	Life, liberty, and security of person Nondiscrimination and equality before the law Access to effective remedy Freedom of opinion, thought, conscience, and religion Freedom of assembly and association Privacy Freedom from arbitrary arrest, detention, and exile; right to a fair trial; innocence,before being proven guilty	- Accuracy in identifying perpetrator-only content - Criteria to assess significance of online presence - Assessment of violent or extremist content - Probability of false positive
Activate and notify: activate the protocol, notify the members, inform them the level of action needed	During the activation stage, the protocol operators will reach out to their members and other stakeholders with information about the attack. As the activate and notify stage provides guidance to various members about the level of action that is needed, it has a potential high impact on human rights. The accuracy of information they have received, the use of OSINT and accuracy of the assessment whether it's a violent, extremist attack or not, will have an impact on human rights. Victims of terrorist activities' right to life, liberty, and security might be violated in case of wrong assessment of the event and lack of action. For example, if live streaming is not interrupted, it might potentially impact the right to life and liberty. If a wrong assessment was made and the material was not violent extremist, it could lead to broadening the scope of the protocol. This could have an impact on human rights defenders that might have their content taken down or users that might have their accounts blocked (right to opinion and freedom of expression). Combined with the assessment stage, this stage has a high impact on human rights.	Life, liberty, and security of person Nondiscrimination and equality before the law Access to effective remedy Freedom of opinion, thought, conscience, and religion Freedom of assembly and association Privacy Freedom from arbitrary arrest, detention, and exile; right to a fair trial; innocence,before being proven guilty	- Accuracy and completeness of guidance

Prepare and Act (information sharing stage): look at OSINT materials, share hashes, share awareness about where the content is, share the outcome, ongoing strategic communications	Victims of terrorism and violent extremism right to life, liberty, and security of person and right to effective remedy might be violated in case of transmitting wrong information that could for example lead to deterioration of the situation. For example, if content is taken down the terrorists might become more violent and kill more people. Or at this stage if content is taken down without preservation then the right to effective remedy might be violated. Victims of efforts to counter terrorism and violent extremism right to freedom of opinion, thought, conscience, and religion, as well as freedom of assembly and association, might be violated due to take-down, deamplification and suspension of accounts.Their right to privacy might also be affected due to use of OSINT (since it can lead to profiling) and hash-sharing. Human rights defenders right to freedom of assembly and association as well as freedom of opinion might also be violated if actions such as hash-sharing lead to take-down and deamplification.	Life, liberty, and security of person, Nondiscrimination and equality before the law Access to effective remedy Freedom of opinion, thought, conscience, and religion Freedom of assembly and association Privacy Freedom from arbitrary arrest, detention, and exile; right to a fair trial; innocence,before being proven guilty	- Expansion of hash database - Accuracy and completeness of information sharing (companies act based on their own policy) - Take down of content - Actions taken other than content take-down Sharing hashes with a stakeholder
Conclude: Assessment against thresh-old, monitoring goes down, and producing sum-maries of what's gone on	The kinds of actions that are taken during this phase can have future human rights implications of operating a crisis protocol. In conclude the operators should pay special attention to how their actions impacted human rights at each stage of crisis protocol and come up with a mitigation plan for future incidents.	Life, liberty, and security of person Nondiscrimination and equali-ty before the law Access to effective remedy Freedom of opinion, thought, conscience, and religion Freedom of assembly and association Privacy Freedom from arbitrary arrest, detention, and exile; right to a fair trial; innocence,before being proven guilty	- Diversity of stakeholder consultation - Mitigation plan with a human rights analysis for future events

GIFCT WORKING GROUPS OUTPUT 2022

Crisis Response Protocols:
Mapping & Gap Analysis
GIFCT Crisis Response Working Group



GIFCT
Global Internet Forum to Counter Terrorism

New Zealand Government Representative

About this project

The Global Internet Forum to Counter Terrorism (GIFCT) has a multi-stakeholder working group dedicated to crisis response and incident protocols. The purpose of the Crisis Response Working Group (CRWG) is to improve members' collective ability to respond to online content incidents arising from real-world terrorist and violent extremist attacks in a manner that respects and protects human rights and a free, open, and secure internet.

In 2021–2022, CRWG has built on its level-setting exercise work in 2020-2021[1] by conducting a more detailed survey and mapping of the crisis response landscape. The aim of this exercise was to ensure that CRWG members' awareness was comprehensive and up-to-date and to inform CRWG's analytical and practical work in strengthening the multi-stakeholder response. This CRWG project also contributes to the Christchurch Call Community's Second Anniversary shared work plan for crisis response,[2] which called for a comprehensive mapping of all protocols that (a) defines the role of each, (b) describes individual thresholds for activation and stakeholder responsibilities, and (c) identifies where there are overlaps and gaps.

The project began with a survey of governments to check whether any had protocols in place or under development of which CRWG was not already aware. The next step was to develop and send a detailed questionnaire to the known protocol owners – the European Commission, GIFCT, the Christchurch Call, Australia, New Zealand, and the United Kingdom. The responses were collated in a detailed mapping and a gap analysis was conducted. CRWG has considered the views of protocol owners and stakeholders and has approached the exercise from both conceptual and thematic angles as well as operational ones, drawing on lessons learned from real-world experiences as well as tabletop exercises.[3]

Mapping the Protocols

CRWG's survey of the crisis response landscape as it exists in 2022 has not brought to light any unidentified protocols among governments in GIFCT Incident Response Directory or the larger set of Christchurch Call-supporting governments.

The United Kingdom's domestic protocol is the oldest. It was developed in 2017 after several terrorist incidents that year with an online dimension, including the Manchester Arena bombing in May. All the other multi-party and domestic protocols were developed in separate processes (but in view of each other) during the second half of 2019 following the launch of the Christchurch Call. All protocol owners (including the United Kingdom) support the Call and have committed to developing processes allowing governments and online service providers to respond rapidly, effectively, and in a coordinated manner to the dissemination of terrorist or violent extremist content arising from a real-world attack.

1 "GIFCT Crisis Response Working Group Annual Output," Global Internet Forum to Counter Terrorism, July 2021, https://gifct.org/wp-content/uploads/2021/07/GIFCT-CrisisWorkingGroup21-AnnualOutput.pdf.

2 See Second Anniversary of the Christchurch Call Summit, Joint Statement by Prime Minister Rt Hon Jacinda Ardern and His Excellency President Emmanuel Macron, co-founders of the Christchurch Call, May 2021, https://www.christchurchcall.com/supporters.html.

3 Note that this report predates the formal debrief on GIFCT's response to the Buffalo shooting and will need to be updated in light of those findings and recommendations.

The earliest and key line of effort to fulfill that commitment was the development of the Call's Crisis Response Protocol (CRP), out of which GIFCT's own Content Incident Protocol (CIP) was built and into which it docked. Google hosted a workshop in Wellington in December 2019, facilitated by the Atlantic Council and involving GIFCT, its member companies, its Independent Advisory Committee members, Call-supporter governments, and civil society experts to test these arrangements and generate recommendations to improve them.

Almost three years have passed, and all the protocols remain – appropriately – works in progress. They are dynamic instruments that are tested, iterated, expanded, and refined based on experience and as real-world threats, technical capabilities, and policy contexts change.

Since 2019, the protocols have been tested on multiple occasions, including in response to shocking and tragic real-world attacks, and as a result have been updated and expanded in different ways. For example, GIFCT has used its experience responding to incidents since 2019 to develop the CIP into a more comprehensive, three-tiered Incident Response Framework (IRF) of which the CIP is the highest level. The Christchurch Call implemented an update to its CRP in 2021. Europol hosted a tabletop exercise involving all multi-party and domestic protocol owners in November 2021. The European Commission, as Chair of the EU Internet Forum, is currently leading work to develop guidance on crisis communications for inclusion in its protocol.

Reflecting their shared origins, the protocols are similar in nature, purpose, aim, scope, and usage. These similarities are useful for interoperability. There are, however, some important differences too.

Nature

All the protocols are voluntary in nature but grounded in robust legal frameworks that ensure due process and protection or respect for human rights. The protocols do not in any way override those legal frameworks at the international, national, or regional level.

Purpose and aims

All the protocols are designed to enable a rapid, coordinated, and effective response to an online content incident or crisis. The government-led protocols do so by enhancing communications among the participants, especially in relation to online service providers. Whether communications are enhanced at the operational or executive level depends on the jurisdiction. The government-led protocols aim to prevent and reduce harm to individuals, communities, and the public, while denying perpetrator(s) the opportunity to amplify their messages, gain notoriety and incite others, and further their cause.

As an industry-led arrangement, the purpose of the GIFCT IRF is necessarily different but complementary. The focus of the GIFCT IRF is facilitating rapid information sharing with and among its member companies. The purpose is to improve situational awareness and, should the CIP be activated, to enable hash sharing so that those of its member companies that allow user-generated content can find and remove content quickly in accordance with their respective policies and procedures.

Scope: Harmfulness of content

All the protocols focus on extreme violent content. They require that the material depict or call for imminent serious harm to life.

The EU and Christchurch Call protocols require the content to be linked to a suspected real-world terrorist or violent extremist attack. The GIFCT protocol also allows for coverage of "mass violence," acknowledging that it can be difficult to establish the link to terrorism in the early stages of an online incident and limitations around its current approach to defining terrorism. The Australian protocol covers terrorist material as well as material that depicts abhorrent and extreme violent conduct[4]. The New Zealand domestic protocol also covers other kinds of significantly harmful material[5].

Scope: Who produced the content?

GIFCT requires that terrorist or mass violent content be perpetrator- or accomplice-produced but excludes bystander footage. Other protocols also tend to focus on perpetrator- or accomplice-produced content, but take a more case-by-case approach to bystander footage. For example, the UK protocol has bystander footage in scope where it exceeds a threshold and breaches online providers' terms of service. The judgment often depends on inferring the purpose of the person in producing and sharing the content; where they are acting in support of the attacker and their cause, the content would be in scope.

Scope: Is it a crisis?

Each protocol has activation criteria related to the nature of the content. Most also have criteria or thresholds for determining whether the situation is a crisis, and they are reasonably well aligned across the protocols. Decision-makers are typically required to assess how fast and widely the content is spreading (or likely to spread), and how many countries and online service providers may be impacted. For example, the EU Crisis Protocol contains a risk matrix which has also been incorporated into the Christchurch Call CRP. Fundamentally, these assessments are about determining whether usual governmental or business processes are adequate to find and refer/remove or otherwise act on the content, or whether enhanced communications and cooperation are necessary.

Usage

As was observed by CRWG in 2021–2022, these protocols are for extraordinary situations. Business for Social Responsibility (BSR) has noted that crisis response entails making decisions at speed and therefore mistakes may be made, impacting human rights or internet freedoms.[6] It is therefore important that activation criteria and

4 See the Subdivision H of Division 474 of the Australian Criminal Code and the Online Safety Act 2021 for a more detailed definition of 'abhorrent violent conduct'.

5 This includes content that is or is likely to be 'objectionable' in New Zealand's Films, Videos, and Publications Classification Act 1993, and/or content that should not be visible and viewed by vulnerable members of society due to the level of harm it can cause.

6 Business for Social Responsibility , "Human Rights Assessment: Global Internet Forum to Counter Terrorism," Global Internet Forum to Counter Terrorism, July, 2021, https://gifct.org/wp-content/uploads/2021/07/BSR_GIFCT_HRIA.pdf.

thresholds are reasonably robust, especially when activating a protocol engages additional powers or tools.

It is good that activations therefore remain rare; for example, GIFCT has only activated a CIP on three occasions, and in general the protocols have not been used to initiate many coordinated content takedowns. Nevertheless, the protocols have been successful in connecting and strengthening the relationships between the relevant players across sectors, increasing situational awareness of online bad actors and violating content, and improving monitoring, coordination, and communications.

Strengthening crisis response

CRWG considered the desired outcomes for crisis response – i.e., what success would look like – as well as the elements that must be in place to achieve that and potential gaps. It is worth noting that none of the gaps were previously unknown to CRWG or the wider crisis response community, although recent events (e.g., in Ukraine and Buffalo) have thrown some into starker relief. The good news is that there is already considerable work underway within GIFCT and elsewhere to address known weaknesses.

The value of the crisis response mapping exercise has been in systematically thinking through and comprehensively laying out the known issues as a basis for CRWG to prioritize its own work, for the Independent Advisory Committee to advise GIFCT on priorities for organizational development, and for the broader crisis response community to move forward together. On that basis, CRWG makes the following recommendations:

Scope

- GIFCT should continue work in 2022 towards a **comprehensive, behavior-based definitional framework** for its work, including the IRF. This is critical in addressing the trend away from attacks by proscribed groups and towards attacks by individuals inspired and motivated by disparate ideologies in online extremist communities. This work may also assist other industry bodies like Tech Against Terrorism and companies inside and outside GIFCT in developing their own more comprehensive definitional frameworks.
- CRWG should convene an expert discussion on **legitimate exclusions and the treatment of bystander footage** in the IRF, other protocols, and GIFCT member companies' terms of service. CRWG's tabletop exercise in April 2022 highlighted the challenges of differentiating content captured from different vantage points (perpetrator, CCTV, bystander) and/or shared for different purposes (e.g., in condemnation or support of the attack, as an eyewitness account, as a safety message, or as part of journalist reporting). A deep dive on this subject would help clarify the boundaries between violative and legitimate uses of content in crisis situations.
- CRWG may also wish to consider the treatment of content related to acts of **state-sponsored terrorism and violent extremism** across the different crisis response protocols, in light of recent events in Afghanistan and Ukraine.

Participation

- GIFCT and the Christchurch Call should extend the Incident Response Directory and Crisis Response

Protocol to more **governments** (subject to appropriate criteria and safeguards). This is a particular priority for the Christchurch Call, as it is the only mechanism available to most non-EU governments to initiate a collective response. CRWG may also convene a discussion for all protocol owners and participants on how best to engage with **third countries** (currently outside the protocol), especially where legal frameworks and human rights protections are less developed.

- CRWG should identify the **different kinds of online services** that may be exploited by terrorists, violent extremists, and supporters during and immediately after an attack and think about how to achieve broader engagement and effective coverage of this online ecosystem. The crisis response community should support **GIFCT's** efforts to identify companies aligned with its mission and to bring them on board as members, using the mentoring services provided by Tech Against Terrorism. We should also support **Tech Against Terrorism** in developing its Terrorist Content Analytics Platform (TCAP) to deal with a broader range of content types and extend its alerting function. Working with GIFCT and Tech Against Terrorism, we should survey a range of **smaller platforms** to understand any barriers they face to mounting effective crisis responses, and what additional shared tooling and other practical supports would be useful. Finally, we should discuss how to deal with those online service providers that resist self-regulation and voluntary cooperation, including through **legislative and enforcement action.**
- CRWG should continue to develop the **role of civil society and researchers** in crisis response. As recognized in the Christchurch Call and the Bergen Plan of Action, there is potential to better utilize the expertise and skills of a global network of individuals and organizations committed to combatting terrorist and violent extremist content online and realizing a vision of the internet as a force for good. For example, they could assist GIFCT companies and governments in finding violative content across the internet and address it quickly in a rights-respecting way.

Operational Issues

- To meet the expectations of stakeholders, GIFCT should maintain its current **capacity to monitor and respond to incidents 24/7** and work with its member companies to build the resiliency of that posture.
- GIFCT should build on successful technical tools such as hashing and matching videos/images and early-stage solutions for text and PDFs to explore solutions to hash **audio content**. GIFCT should also work with member companies to fully operationalize the hashing of **URLs** in the TCAP as a powerful way of disrupting out linking from GIFCT members' platforms to off-platform repositories of CIP-related content.

Data Preservation and Access

- Building on CRWG's report in 2020–21 and Europol's November 2021 tabletop exercise, the crisis response community should find appropriate ways to advance the discussion of the principles guiding proactive information sharing in threat-to-life situations, and proactive data preservation for domestic and cross-border law enforcement purposes, with a view to making concrete progress in these areas.
- Both the Christchurch Call and EU protocols point to the desirability of preserving data so it can also be accessed for other legitimate purposes, including journalism, research, international investigations, and judicial processes. This is key to operationalizing victims' right to access effective remedies. GIFCT and the wider crisis response community should continue to explore and further develop solutions like Tech Against Terrorism's TCAP archive and integrate them into a coordinated and comprehensive crisis response system.

Crisis Communications

- According to the EU's Radicalisation Awareness Network (RAN) Policy Support, good crisis communications can reduce the opportunity for terrorists and violent extremists to draw strategic benefits in the aftermath of an attack. The Christchurch Call and other protocol owners should therefore consider developing strategic communications frameworks, building on the principles in the EU Crisis Protocol and the work being done by RAN-PS on practical guidance for implementing the principles.

Cross-Cutting Issues

- GIFCT and other protocol owners should apply the human rights matrix developed in CRWG in 2021–22. The purpose of the matrix is to help practitioners identify the individuals and groups whose rights may be at risk at different points in the "lifecycle" of a response with a view to preventing or mitigating negative impacts. The matrix is a work in progress, and in the future CRWG could focus on refining the human rights indicators in the matrix and developing a framework for assessing the impacts of actions at each stage of response. Protocol owners may wish to tailor the matrix and indicators to their process and practice using them in exercises. Stakeholders may also use this work to structure a human rights impact assessment as part of any debrief and review.

Immediate next steps

- GIFCT will provide a formal debrief of the Buffalo incident in accordance with the framework developed by CRWG last year.[7] That will be an opportunity for the organization, its member companies, and government and civil society stakeholders to reflect on the incident and the response, and to identify ways to improve again on our collective response networks, tools, and processes. As a contribution to future debriefs, and as part of ongoing work to develop good practices for transparency around the different stages of crisis response, CRWG could do work in 2022–23 on the best **metrics for evaluation**, so we can better assess our progress and demonstrate it to others.
- CRWG will use this mapping and gap analysis, complemented by the Buffalo debrief findings and recommendations, as a basis for identifying and prioritizing its work in 2022–23 and beyond. CRWG will ask protocol owners to update the information in the mapping each year to ensure members' knowledge of the crisis response landscape remains up-to-date and as a basis for more detailed investigations in specific areas.

7 "GIFCT Crisis Response Working Group Annual Output," Global Internet Forum to Counter Terrorism, July 2021, https://gifct.org/wp-content/uploads/2021/07/GIFCT-CrisisWorkingGroup21-AnnualOutput.pdf.

GIFCT WORKING GROUPS OUTPUT 2022

Crisis Response & Incident Protocols 2022 Tabletop Exercise Public Report

GIFCT Crisis Response Working Group

GIFCT
Global Internet Forum to Counter Terrorism

Exercise Descriptions & Purpose

A. In order to perform an extensive and productive assessment, the team used tabletop exercises designed particularly for different stages of GIFCT's workload. The tabletop exercises were developed to simulate crisis situations to review roles and emergency responses of GIFCT, member companies, and other stakeholders.

B. During each exercise, the design team from KizBasina (NGO Sector Facilitator) was available to ensure smooth operation of the exercises.

C. For the purpose of easy introduction to first-time participants, each exercise was designed using different scenarios, visual and written aids, and charts.

D. For the assessment, three exercises were planned:

- **TTX1 - Human Rights Exercise:** The exercise concentrates on the impact of the GIFCT Incident Response Framework on Human Rights and how to ensure that they are appropriately balanced and protected.
- **TTX2 - Communications Exercise:** The exercise tests the efficacy and quality of current processes and procedures between GIFCT team and members as they navigate the Incident Response Framework.
- **TTX3 - STRATCOM Exercise:** The aim of the exercise is to test the efficacy of GIFCT's public statements to serve their intended purpose during an incident and to assess the quality of the messaging included in GIFCT's public statements.

E. Unfortunately, the TTX3 exercise (STRATCOM) was not able to be performed at the designated time due to extraordinary circumstances. It was scheduled the week following the attacks in Buffalo, NY and so participants were fully engaged managing the active crisis. GIFCT released public statements on both the activation of their Content Incident Protocol in response to the attack[1] and their debrief process, which contain further lessons learned relevant to this report.[2]

F. This report represents a summary of the findings from the exercises; more detailed conclusions have also been provided to GIFCT and the Crisis Response Working Group. However, the public release of these details would provide insight into the incident response processes of GIFCT and their partners, which may provide terrorists and violent extremists with information that aids in their adversarial behavior.

Lessons Learned - Design and Facilitation of Exercises

During and following TTX1 and TTX2, the project team identified the following lessons in order to maintain an accurate simulation for the participants.

1. **The importance of player aids** is one of the most important parts of the tabletop exercise. While they are designed to be as simple and easy as possible, without explanations or additional material to guide the players, the exercises easily present complications that may disturb the simulation. The team found it

1 Update: Content Incident Protocol Activated in Response to Shooting in Buffalo, New York United States, GIFCT, May 18, 2022, https://gifct.org/2022/05/14/cip-activated-buffalo-new-york-shooting/.

2 GIFCT. (2022b, June 23). Debrief: CIP Activation, Buffalo, New York USA. https://gifct.org/2022/06/23/debrief-cip-activation-buffalo/

best when different aids were used to explain the purpose, timeline, and stages of the exercise.

2. **The number of participants** involved during the exercise is also another variable that can change the course of the facilitation of the exercise. As participants or groups increase, the number of facilitators should proportionally multiply in order to respond to the needs and questions of the participants as effectively and rapidly as possible.
3. **Lastly, as online communication methods**, direct video-sound based applications have been shown to be more valuable in view of the fact that giving and receiving instantaneous information is easier. As opposed to messaging/community apps, they allow a more clear, open, and instant communication channel among participants and the facilitation team where acknowledgment of the transmitted information is instant and assured. However, it is important to note that for larger groups of participants (30+) they are not as useful and can be complicated. A balance of these tools should be considered for both effective exercises and effective crisis response.

Lessons Learned - Protocol Design and Operational Response

In consideration of the purposes and outcomes of the exercises and upon further assessment of participants' actions, the project team came to the following conclusions:

A. **Clearer definitions of certain terms employed in the guidelines:** More clear and understandable definitions of key terms placed in the protocols would increase response time, reduce confusion, and mitigate the risk of civil freedom concerns.

B. **Clearer guidelines on event assessment**: In crisis situations where time is of the essence for GIFCT, member companies, and government stakeholders to fully and correctly assess an incident, a minimum definition or a guide to characterizing an event is important.

C. **More stakeholder discussions:** As GIFCT operates on a case-by-case basis, with every new incident a new problem or obstacle may arise. In consideration of the large number of stakeholders GIFCT works with, it is important that they keep up with current news and trends. Monthly events where experts speak on current issues with the participation of GIFCT stakeholder representatives may be a solution.

D. **More comprehensive protocols:** Current protocols should be definable, defensible, and scalable across situations. They need to be updated and extended regularly with every simulation or real-life incident, and should be clear not only on definitions but paths to follow in crisis situations. Protocols should be clear about the expectations of each stakeholder group participating, laying out responsibilities as well as what they can expect from the protocol.

E. **GIFCT communication:** During crisis response, communications must be timely, simple and clear - especially in writing. Communications templates should be enhanced to ensure efficiency and accessibility while also removing unnecessary duplication. Tech companies and other stakeholders should also consider standardizing communications to ensure clear, concise, timely, and complete communication.

Additional suggestions from stakeholders were:

- Providing additional methods for stakeholders to raise concerns or issues during exceptional circumstances;
- Enhancing expectation settings about harm types in scope and developing proactive

communication; and

- Increasing transparency by using sanitized versions of the internal debriefs during public communications.

F. **Existing communication channels:** Current communication mediums can be improved, especially considering instant response and acknowledgment features. Furthermore, two-factor authentication and the importance of fallback systems and secondary communication channels in the case of a malfunction were also viewed as critical.

GIFCT WORKING GROUPS OUTPUT 2022

Active Strategic Communications: Measuring Impact and Audience Engagement

GIFCT Positive Interventions and Strategic Communications Working Group

GIFCT
Global Internet Forum to Counter Terrorism

Munir Zamir
University of South Wales

Introduction

The Global Internet Forum to Counter Terrorism (GIFCT) aims to prevent the proliferation and promotion of terrorist and violent extremist basic content on the internet through collaboration between the tech sector, civil society and governments. Through this strategic and operational partnership, GIFCT works with practitioners, academics and agencies to create work streams and paths to help inform, identify and tackle the multi-layered problem sets presented by harmful actors in the online space. Through its Positive Interventions Working Group (PIWG), an output related to better understanding of active strategic communication from measurement, impact and audience contexts was agreed. The following report provides key learning, barriers and challenges to these issues, as a means to progress this subject area into a more evidence based arena for future efforts.
This report focuses its attention on the processes, practices, and challenges of designing, delivering, and measuring online positive interventions within Countering Violent Extremism (CVE) and Counter-Terrorism operational contexts. As such, the report aims to continue (in a developmental manner) the building and sharing of knowledge, practices, and learning that has been led by GIFCT's CAPPI WG on positive interventions. Building on the July 2021 CAPI2 "Positive Interventions" report, which provided a macro-level view of strategy, delivery, and program considerations, this report outlines a more granular and practitioner-oriented effort, focusing on specific elements of the strategic communications process and how this affects and impacts understanding of audience engagement, reach, and measurement. Such an approach offers insights and learning from localized, global, and private-public partnerships that have been created specifically (or in alignment with) CVE needs such as counter-disinformation, harm reduction, critical thinking, and prebunking/inoculation-based prevention work.

There are three main sections of this report that delve into the outputs the CAPPI WG identified as areas of interest: measuring impact or success in campaigns, best practices for audience targeting, and turning passive counter-narratives into active strategic communications. The core content of the report provides examples, conceptual threads, and suggestions for how to make strategic communications more active and grounded through nuanced understandings of audiences, sentiment, and engagement. In doing so, the overall narrative emerging from this effort points to the need for private-public partnerships to evolve in order that they sit firmly at the heart of CVE efforts, driving innovation, trust building, and impact measurements moving forward.

Section 1: Measuring Impact or Success of Campaigns

In the online campaign or project delivery context, the need for measurement and evaluation systems to be part and parcel of the overall design is an accepted norm and practice. Within the commercial arena, this often takes the shape of a set of impact indicators that have become widely known as 'vanity' metrics. The main aim of such indicators is to provide both the delivery agent (advertising agency, government agency/Civil Society Organization (CSO)/NGO) and the 'client' or principal stakeholder with a set of measurements (likes, shares, impressions) and the number of times a piece of content is 'seen' and then 'engaged' further with – known as a Click Through Rate (CTR) – that offer a birds-eye view of the project's outcomes. These metrics also serve as barometers for overall campaign/project success and value for money discussions. However, this practice requires further development with regards to how such measurements can be viewed within more detailed qualitative and quantitative aspects of behavior change efforts, including greater emphasis on capturing sentiment, longitudinal impact, and the role of other interlocutors (for example the individual's own offline/online eco-system). Within the CVE space, this issue and set of practices take on greater degrees of nuance in regard

to how accurately and effectively such metrics can offer meaningful data and results for highly subjective and open-ended issues such as radicalization, deradicalization, disengagement, and desistence. The idea of online positive interventions posits certain assumptions about both the role and potential attitudinal and behavioral impact of the intervention on often hard-to-reach audiences with diverse levels of vulnerability and disparities in their access to information.

Based on GIFCT's CAPPI WG CAPI2 report published in 2021, positive online interventions are operating under the rational belief of the effectiveness of promoting credible, positive alternatives or counter-narratives, while acknowledging that end-users are not simply passive recipients of messaging and the values contained in them. This forms the 'what' element of the project's inception. The second element takes on the need for an overarching goal or objective, which is to counteract possible interest in terrorist or violent extremist groups. This goal also operates under the caveat that interaction and engagement with ideas put forward by positive interventions would be effective and leave an impressive legacy as a result of the experience. Such goals become the 'why' answer for the intervention's need to exist or be initiated. In a related effort, work carried out under the European Commission's Radicalization Awareness Network (RAN) Communications and Narrative WG has initiated and continued to develop the GAMMMA+ model as a practical guide tool for conducting positive interventions within an overarching framework model.[1]

A United Nations Development Programme (UNDP) and International Alert collaboration produced a report based on an assessment of a CSO project by PeaceGeeks that lent its focus to different types of proxy variables and attitude types that were possibly indicative of extremist beliefs. Such efforts are seeking to build out and develop a more robust understanding of both the identification and prevention of extremist attitudes.[2]

Wrapped up within theoretical framework contexts, questions of measurement become aligned to the idea of impact, with the issue of "reach" being viewed as a standard cumulative metric that can offer both evaluative and success criteria outputs for project performance. While the idea of offering reach statistics appears to be logical in terms of wanting to know to who, where, and even how the message, content, or communication was presented, reach can lead to misleading and counter-intuitive sets of metrics without certain key "add-on" features incorporated into the data mining mix. It is even debatable if reach alone can be viewed as equating to impact, and if CVE campaigns should even be aiming for optimum or maximum reach, given the contested nature of radicalization as a concept and the dangers of mass audience targeting in this context. In essence, when assessing reach metrics in the CVE online interventions space, certain unique factors and conceptual issues affect project performance in ways that can skew both what a project has set out to achieve and the results that are gathered from vanity metrics and reach measurements. These warrant further discussion in order to ensure a more effective relationship between project design and project outcome. The key to online CVE interventions rests with the planning and insight phase, as these two issues directly affect audience targeting, project measurement, and performance results, as well as creating a baseline assessment through which impact is ultimately understood and evidenced later on.

In order for CVE online interventions to move beyond reach statistics, the design of such efforts should prioritize

1 "RAN C&N Effective Narratives: Updating the GAMMMA+ model, Brussels 14-15 November 2019," European Commission: Migration and Home Affairs, n.d., https://ec.europa.eu/home-affairs/pages/page/ran-cn-effective-narratives-updating-gammma-model-brussels-14-15-november-2019_en.

2 "Design, Monitoring & Evaluation of PVE Projects in Jordan - A Baseline Assessment of PeaceGeeks' Projects," International Alert, 2022, http://www.pvetoolkit.org/design-monitoring-pve-project-in-jordan.

incentivizing engagement upon being reached instead of merely reach statistics. What matters is the nature of the engagement and it yielding positive outcomes for deradicalization /disengagement efforts. Reach data alone can be too broad to provide specificity in counter-radicalization efforts. For example, one of the first questions that needs to be addressed is what the evaluation needs actually are? Evaluating projects in order to prove the efficacy of a policy stance, such as the UK government's efforts to counter extremism through its "Prevent" agenda, has been seen as a "scatter-shot" approach designed to produce 'quick wins' because of a lack of perceived risk and proportionality rationale. Evaluations can also be conducted to justify the project's existence beyond the 'why' logic framework, but these typically rely on highly subjective and contested ideas such as values, citizenship, and certain definitions of what 'moderate' versus 'extreme' views may mean.

This is evident in the UK government's social cohesion and counter-extremism approach ("Contest Strategy"), which has often conflated promoting cohesion to mean the reduction of extremism by using behavioral science-based evaluations in such a broad context as to undermine producing reliable data/outcomes. Although this example is from a policy perspective, the operational issues of conflating 'mainstream' social cohesion efforts with countering extremism often stem from having too broad an approach to defining problem sets and targeting audiences without adequate insight or rationale. The issues of seeking to increase cohesion levels and reducing the influence of extremism have now started to be separated through more distinct lines of effort. For example, GIFCT's theoretical framework approach (mentioned above) differentiates between resilience building within a prevention context (further from harm) and undermining extremist ideology in more upstream counter-extremism contexts (further towards harm). By way of example, the UK Home Office's "Building a Stronger Britain Together" initiative via the work of its Research Information and Communications Unit (RICU) offers detailed insight into these issues.[3]

An even more pertinent strand of this issue would be to ascertain if the evaluation is needed in order to measure or present an empirical analysis of successful deradicalization or disengagement. In this context, the type of campaign and its tactics (e.g., resilience building, capacity building, support services, deradicalization, or counter-radicalization) should ultimately determine the nature and role of the evaluation approach as well as its suitability in achieving aims and objectives.[4] There is little doubt that reach statistics need to provide many stakeholders and interested parties with 'catch-all' data sets that can serve different purposes depending on the line of inquiry. Both offline and online interventions in this space are unable to claim any sole causal link to someone's life journey, choices, and influences over time, as they are based purely on the 'impact' of the intervention. As such, no measurement and evaluation model or framework can make any similar claim. Therefore, the goal of such work is based on the mitigation and prevention of harm at the deterrence level alongside challenging and exposing harmful narratives at the undermining level of such interventions. Such efforts are designed to reduce the likelihood of harm and seek to protect those who may be most vulnerable to narratives and potential actions. The "measurement of success debate" is one that should be seen as working towards creating better practices through a greater understanding of what works and what doesn't in specific contexts and environments. The better the specific measurable outcomes, the greater the likelihood of the overall online CVE interventions space being more and more effective.

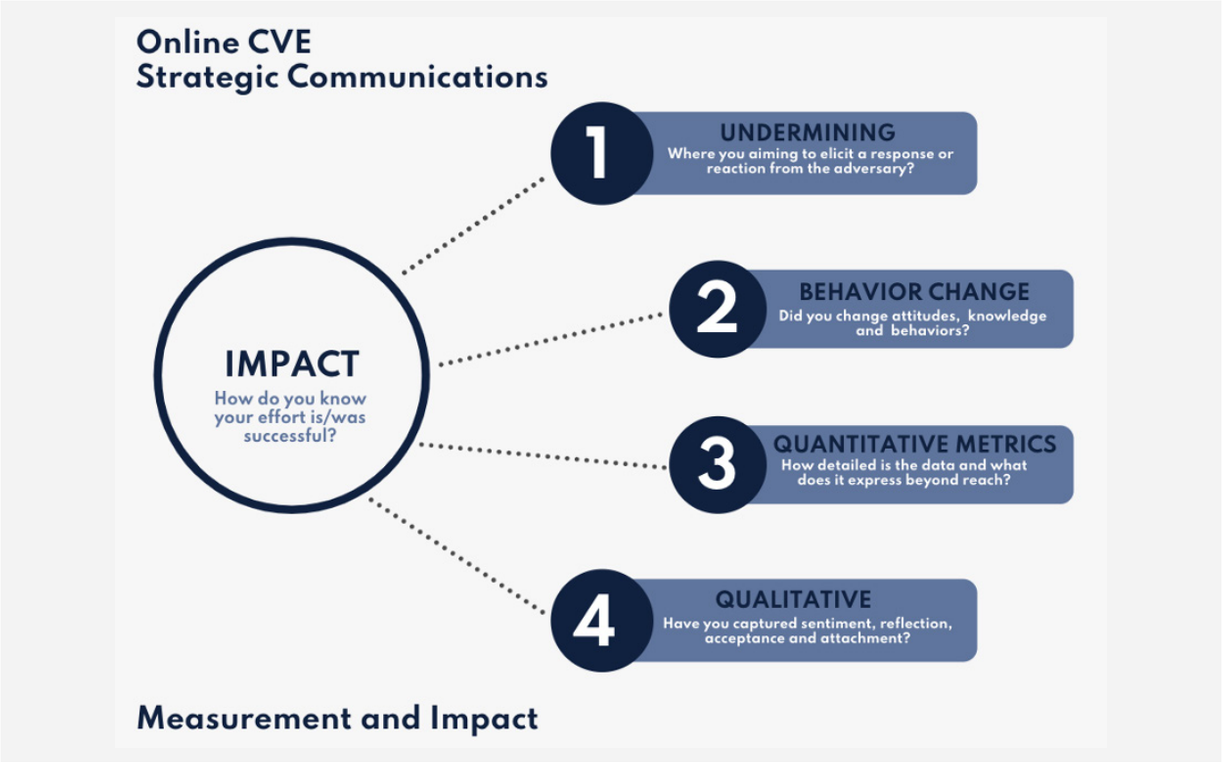
3 Home Office, "Evaluation of the Building a Stronger Britain Together (BSBT) programme," GOV.UK, July 29, 2021, https://www.gov.uk/government/publications/evaluation-of-the-building-a-safer-britain-together-bsbt-programme.

4 "CTED Analytical Brief Countering Terrorist Narratives Online and Offline," United Nations: Security Council - Counter-Terrorism Committee (CTC), 2020, https://www.un.org/securitycouncil/ctc/content/cted-analytical-brief-%E2%80%93-countering-terrorist-narratives-online-and-offline.

The aim of isolating and identifying true causal links to certain decisions or actions/ behaviors is complicated by the fact that online identity is couched inside numerous "influence" parameters that are impossible to capture and codify. This is even more nuanced in the realm of CVE and it is of great importance that the strategic framework created for an online intervention must offer an exceptional level of understanding of the problem set, insight into audience types and lifestyles, as well as key expertise in the problem set being explored (be it deradicalization or capacity/resilience building). Simply knowing the type of intervention required at the strategic level (deter, redirect, etc.) does not address the depth of insight needed to deliver the intervention with an effective plan and model. This applies equally to the issues of reach and ensuring the "right" audience(s) have been reached. This is where the strategy and design phase are so crucial in regards to the direct relationship that exists between insight and audience – namely the more robust the insight, the more accurate the audience targeting.

While it is acknowledged that the size, scale, and context of the protagonists involved in this space create a huge variance in regard to capacity and resourcing matters, it must also be acknowledged that good planning and strategy will fundamentally improve the prospects of a campaign being "successful" or "effective." Therefore, certain useful practices can be put in place (e.g., top-level strategy, theory of change, effects framework, adversary analysis, audience segmentation matrix) to ensure that design, delivery, and measurement are not only aligned with each other but also with the overall objective of the project or campaign in question. Some of the key considerations in regards to reach and impact can be addressed by ensuring that a clear set of objectives, indicators, and data collection processes are in place, as per below:

Figure 1: Measurement and Impact Considerations



Strategy and Objectives

Questions and considerations relating to impact and reach are best managed through the development of an effective strategy. Does the strategy and desired end-state align with the objective(s) set out by the stakeholders, problem set, or client? In this context, the term strategy refers to the logical framework, proposition, and rationale created to express what needs to be done, why it needs to be done, and how it will be done. A key area that is often overlooked in CVE online intervention planning is the issue of the effects (both intended and unintended) of the campaign, message, or content being part of the public domain. Strategy and its efficacy form the core component of what needs to be done and how this will be done:

Figure 2: Strategy Considerations

A practice that can to some extent help to alleviate over-reliance on generic reach data is the use of an "effects framework," which is essentially a set of outcome indicators at a granular level that can identify and measure the impact of key tactical methods, content, or messaging deployed by a campaign as part of an overall theory of change. As the effects framework needs to take shape during the strategy and planning phase, its outcomes are tied to both strategic intent, objectives, and impact. This comes down to a simple proposition: if, for example, behavior change is one of the key aims of a project, one of the objectives must be something similar to "reducing the impact" or "undermining the narrative" of the antagonist/adversary. The reason for this assertion is obvious in the sense that any intended target audience would need to change its behavior towards a Violent Extremist Organization (VEO) through actions, which would then need to be picked up by the effects framework. This change (ideally) needs to have been planned for, which is why it is part of the effects framework, and as a result should also be able to be identified, measured, or verified. For an online intervention, reach statistics alone cannot sufficiently determine this change or variance, so more nuanced impact and effects measurements are required. The role of strategy in communications in this regard cannot be underestimated.[5]

5 Haseeb Tariq, "Five Components Of A Successful Strategic Communications Plan," Forbes, June 22, 2021, https://www.forbes.com/sites/forbescommunicationscouncil/2021/06/22/five-components-of-a-successful-strategic-communications-plan/?sh=326789e05813.

Going Beyond Reach

Classically, calls to action were seen as useful indicators in assessing the extent to which audiences engaged with a message, understood its premise, and then actioned desirable efforts through active choice and decision making. Advances in social media platforms and tech tools combined with greater degrees of audience lifestyle segmentation have meant that relying solely on online reach data can create misleading and skewed notions of both audiences and their needs. A simple example would be to view an individual's reaction to a piece of content (like, dislike, share, retweet, etc.) as a barometer of their values set, preferences, or position on a subject matter. This also applies to contextual disparities that exist, because one piece of content delivered a certain way may have entirely different sentiment values attached to it by the same person in different circumstances. Such issues are pertinent not just to reach and impact matters, but equally to ensure the differences between extremism and violent extremism.

The "golden goose" of online CVE interventions remains the relationship between content and engagement, followed by values alignment and behavior change. The idea that genuine and often challenging engagement lives beyond generic reach metrics should drive the design and narrative elements of the positive intervention. The notion of active communication is one that essentially elicits a response beyond pressing 'like' and takes both the content and its end-user into a new space of engagement that may be challenging or affirming but allows the end-user to express ideas and sentiments and (ideally) make positive choices. Of course, the context of the campaign largely dictates the type of engagement that may be possible (e.g., building followers and generic metrics versus supporting a local network to improve governance structures to undermine extremist influences), as deradicalization efforts are markedly different from resilience building efforts in terms of tone, message, and objectives as well as the type of action or desirable behavior sought. However, engagement is also important in and of itself, because comments and other types of interactions give other users social clues about the content.[6] Comments can actually change users' perceptions of the video/article they are seeing, so community engagement and management can not only bring important benefits for evaluation[7] but actually support the impact of campaigns.[8]

An area of effort that is gathering pace in regards to both approach and content design is "prebunking"-based interventions. Also known as work that takes an inoculation theory approach as its central proposition, moving firmly into the preventative space with interventions that see "prevention, not cure" as the best way to engage in both the CVE and disinformation/misinformation space, prebunking efforts have made significant strides in both design and capability. A central development in this form of approach has been to move away from the "debunking" tactic into the building of "mental armor" through prebunking content that is designed to prepare someone to identify, assess, and make informed choices when extremist or disinformation content appears on their screen or social media feed. However, there are challenges with this approach, specifically how adding significant size and scale would be managed and still retain elements of nuance and stratification. Future development of this approach can seek to further refine this element. The work of Kurt Braddock provides an excellent roadmap into the emergence of

6 Franklin Waddell, "What Does the Crowd Think?" New Media & Society 20, no. 8 (August 2018): 3068–83. https://doi.org/10.1177%2F1461444817742905.

7 Jae Eun Chung, "Peer Influence of Online Comments in Newspapers: Applying Social Norms and the Social Identification Model of Deindividuation Effects (SIDE)," Social Science Computer Review 37, no. 4 (August 2019): 551–67. https://doi.org/10.1177/0894439318779000.

8 Hue Dong, Hong Vu, and Long TV Nguyen, "Effects of Online Comments on Risk Perception," The 70th Annual Conference of International Communication Association, Gold Coast, Australia, January, 2020. https://www.researchgate.net/publication/338751162_Effects_of_Online_Comments_on_Risk_Perception_and_Intention_to_Communicate.

prebunking/ inoculation efforts into the CVE terrain.[9] In collaboration with Kurt Braddock and others, Jigsaw (a division of Google) has attempted to advance the practical application of debunking efforts and offered both testing phase results as well as setting out more contextual criteria for how this work can continue to develop.[10]

Debunking-based efforts can sometimes struggle to match both the quantity and propensity of extremist content simply because their dissemination does not have a pace sufficient enough to challenge and then move end-users' attention span away from the original harmful content. Some remnants, ideas, or narrative elements remain, and this is why the building of mental, emotional, and digital resilience through prebunking is seen as a more effective method in preventative spaces. The logic behind this approach is simply that if enough examples of extremist-based content are shown to users through prebunking tactics, they will be better equipped to identify and question it.[11]

Measurement and Evaluation

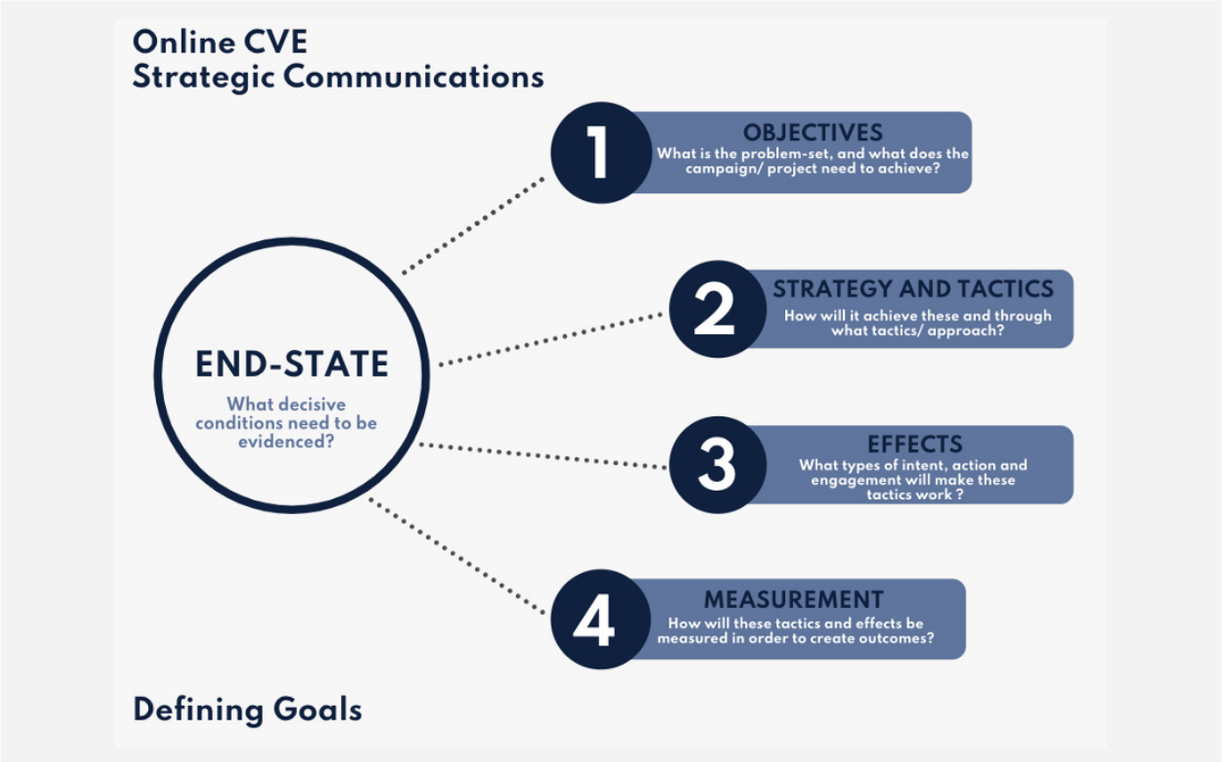
In order to enhance and improve the quality of outcome indicators that a project may be seeking to identify through its content dissemination, the assessment framework requires two core pieces of information at the outset: the desired end-state and the means to have even a simple baseline understanding of the intended audience's level of support, sentiment, grievance, hope, fear, etc., of the issue being tackled. The baseline assessment provides the most robust and reliable method of identifying how much movement is needed during the lifespan of the campaign from the audience and how this can be effectively measured (which should already be incorporated into an effects framework). This baseline assessment also allows the desired end-state to be measured so as to ascertain how well its conditions have been met. If such processes and practices are not in place, there is a strong likelihood that results will not go beyond vanity metrics and birds-eye view statistics that still plague the online CVE space. This highlights the need to have a clear end-state goal in mind in order to be able to measure outcomes and impact appropriately:

9 Kurt Braddock, "Vaccinating Against Hate: Using Attitudinal Inoculation to Confer Resistance to Persuasion by Extremist Propaganda," Terrorism and Political Violence 34, no. 2 (2022): 240-262, doi: 10.1080/09546553.2019.1693370.

10 Kurt Braddock et al., "Engagement in subversive online activity predicts susceptibility to persuasion by far-right extremist propaganda," New Media & Society, (February 2022), https://doi.org/10.1177%2F14614448221077286.

11 "Psychological Inoculation: New Techniques for Fighting Online Extremism," Jigsaw, June 24, 2021, https://medium.com/jigsaw/psychological-inoculation-new-techniques-for-fighting-online-extremism-b156e439af23.

Figure 3: Defining Goals



Sentiment and Engagement

The ideal scenario within an online CVE intervention is that it has some synergy with efforts in the offline space. The idea of an "audience" obscures the reality of terms like 'network,' 'community,' and 'activism' at local levels. Online audiences are mysterious and hard to "read" in the isolation of an offline reality, which exacerbates the issue of some audiences being "hard-to-reach." A way around this quandary is to attempt to base much of a project's content focus on engagement with people offline before they become audiences. However, how much or how little offline engagement may or may not be required is dependent on the campaign's goals and tactics. This approach is also linked to a need for the CVE efforts in the online space to move beyond the traditional focus group route and start to incorporate audiences throughout the inception, dissemination, and evaluation stages. Although resources for CSOs/NGOs are limited, these entities technically still have better access to community networks and actors who help bridge the gap between the project and the experience of the local population. Engagement built from open and trusting practice has a better chance of succeeding than unverifiable campaigns that can arouse suspicion. The closer the offline engagement is to the online iteration, the greater the chance of success for the project over time. This is also true of sentiment and tracking sentiment change. The need to understand the concept of the "audience" in the online space cannot be understated and is very much at the heart of successful planning for sensitive campaigns.[12]

12 Molly Riddle-Nunn, "Tips and Tools for Understanding Your Online Audience," Mostly Serious, June 22, 2018, https://www.mostlyserious.io/news-updates/understanding-your-online-audience.

Meeting Communities Offline

In ideal circumstances, the measurement and impact needs of a project or campaign benefit greatly from practitioners having access and engagement with audiences/ communities in offline settings. In this context a key advantage to delivering positive interventions from a community/CSO/NGO perspective is that of physical access to various potential target audiences. By the nature of their community origins and placement, such organizations are in a position to engage in a more hands-on fashion with networks, interest groups, and individuals to advance project aims and objectives. Traditionally, this means that levels of trust and mutual support offer projects the chance to be more grounded in real-world needs and also allow for greater flexibility in using creative and diverse means and mediums to deliver targets. The approach used by PeaceGeeks (https://peacegeeks.org) to this issue is a very useful place to begin understanding the offline/online nexus. Community engagement that is grounded in offline relationships allows design and planning to go beyond the norm of "focus group" testing with relatively unknown audiences to a process that can add diversity to different possible target audience needs with bespoke roundtable and workshop style events that allow greater nuance and depth to be added to project design elements. This results in delivering content that is to some extent the product of a partnership effort between the CSO/NGO and the community, with the latter taking on the role of stakeholder in a basic form. This type of community engagement can add significant value to the intervention's credibility and longevity among online audiences. Although the CSO and private sector do not always naturally fit together in certain contexts where nuances may need to go beyond "cultural expertise" from a distance and instead be grounded in the real-world experience of vulnerable communities, in terms of the need for efficacy in approach and tactics, the commercial sector possesses several useful avenues for CSOs to explore to enhance their existing offline engagement.[13]

Testing Reactions to Content

A project's chances of creating a successful impact are closely linked to how effectively the testing or pilot phase is at identifying strengths, weaknesses, challenges, and nuances. The initial phase of such testing often involves focus group discussions on a basic concept or idea, which allows both prospective audience members and the project team to understand positive and risky or more negative elements of the design. This can be further enhanced through simple variations to the 'A-B' testing model, such as rendering a few different versions of the same core content and disseminating them to similar target audiences to test and gauge reactions to the subtle variations in content and its presentation. The Abdullah-X project, which was a CSO/YouTube partnership, used A-B testing through creating different thumbnails with varying degrees of intrigue and then targeted different potential users through hashtag/metadata changes with the same core content.[14] A-B testing clearly offers results and data on what content elicits greater engagement, but this is not as easily applied to impact measurement. Another way to elicit more nuanced testing phase feedback is through an iterative approach that is open to comments and suggestions emanating from the sample audience in regards to possible language and tone as well as cultural and related nuances.

There is an argument that testing using more extreme or polarizing content can exacerbate existing conditions.

13 Aritrya Sen, "7 Offline Customer Engagement Strategies You Didn't Know About," Involve Me, September 14, 2020, https://www.involve.me/blog/offline-customer-engagement-strategies/.

14 Abdullah-X, "Freedom of Speech vs. Responsibility," YouTube.com, March, 2014, https://www.youtube.com/user/abdullahx.

Despite this potential risk, this approach also serves as a means to maintain an element of tolerance towards changes needed to content in order to build or maintain user engagement and trust. Such an approach offers what is known as a 'network effect' to come into play; audiences that are familiar with the content, having engaged and been part of a testing phase, can subsequently amplify the reach of the end-content to their online and offline networks.

The Abdullah-X project underwent rigorous testing, feedback, and adjustments from all perspectives, including branding, content style, duration, narrative, sentiment analysis, and audience targeting. The project also engaged with potential target audiences offline, taking both the content and concept into classrooms across the UK, enabling audiences to help shape the overall design and glean information on elements that offered more engagement and curiosity. This process employed A-B testing, where more counter-narrative/challenging versions of content were tested for different thumbnails, hashtags, and metadata to ascertain if more radicalized individuals or ISIS supporters were more or less likely to click depending on what was used.

Translating Offline Efforts into Online Domains

A key way in which offline efforts can form an important aspect of online intervention approaches is through adaptive delivery, which is different from reactive delivery by virtue of its content tone and form. In the adaptive context, project content design retains an element of scope to re-render and re-order narrative tone and form to more closely match the needs of target audiences through engagement analysis. Where content is based on social media posts or audio programming (for example radio/ podcasting), the tolerance for amendments and subtle changes is clearly easier. For more creative delivery such as video, animation, serials, etc., the key learning from the offline domain is to ensure that the initial testing phase can glean feedback on both concept, creative treatment, narrative, and tone.[15]

A key aspect of the offline experience is the individual user experience that can still be found within a group or community delivery context. This translates into the online setting through content that can house intellectual, emotional, and personal "touch points" from one overarching delivery strand. Examples where this method is implied within the project parameters can be found in the One2One[16] campaign and search redirect intervention approach.[17]

When a campaign is seeking to engage mass audiences, narrative form and content style play a large part in either being able to create effects at the individual level through critical reflection or using persuasive messaging to encourage individual actions that help achieve bigger end-state objectives. Both of these types of effects are in their way linked to the idea of content being able to access and influence the cognitive space of the end-user, seeking what is commonly known as a "cognitive opening."[18]

15 Andrew Glazzard, "Losing the Plot: Narrative, Counter-Narrative and Violent Extremism," ICCT, May 22, 2017, https://icct.nl/publication/losing-the-plot-narrative-counter-narrative-and-violent-extremism/.

16 Ross Frenett and Moli Dow, "One to One Online Interventions – A Pilot CVE Methodology," ISD, September, 2015, https://www.isdglobal.org/isd-publications/one-to-one-online-interventions-a-pilot-cve-methodology/.

17 " Redirect Method Canada: Final Report," Moonshot CVE, March, 2021, https://moonshotteam.com/resource/canada-redirect-final-report/.

18 First introduced in Quintan Wiktorowicz, Radical Islam Rising: Muslim Extremism in the West, (Lanham, Md: Rowman and Littlefield Publishers, Inc, 2005).

Another aspect of translating offline methods to the online domain is found in the management of negative feedback, pushback, or "blowback" of content from audiences who may or may not be part of an intended target audience. Often seen as a gray area of the CVE world in terms of a campaign's remit or scope to address such issues, this scenario is often found where campaigns gain the type of trust or credibility usually reserved for face-to-face offline interventions in the deradicalization and prevention space. Dependent on whether a campaign's focus is providing alternative narratives or some form of counter-narrative, negative feedback allows a window of opportunity to directly address grievances, misinformation, or hate speech through tact, poise, and intellectual prowess.

These are attributes not usually associated with extremist narratives, which are generally known to be binary and obstructive in nature and tone. This does not, however, mean that extremist narrative is not complex in narrative, form, and tone. The issue is that pushback means the campaign's overarching presence has caused a reaction from those who are clearly part of the problem. An appropriate well-crafted response is a very effective way to undermine this feedback and build presence in an already congested online space. The opposite applies to rushed, panicked, and not well thought out responses, which may inadvertently amplify the extremist messaging. Ultimately, the scope and context of a project will determine if this approach is viable or not, but the fact remains that direct engagement of negative feedback that is well-constructed and tactical in nature can provide a campaign with enhanced trust and behavior change possibilities. Both audience feedback and behavior change issues fall neatly into the next section, which looks at how engagement is understood through audience sentiment broadly.

Different Approaches to Measuring Sentiment Change

The role and scope of communications-based efforts to counter and challenge issues like extremism remain limited without other inputs being in place at a larger scale. While this can be construed as a loss-leader if taken literally, the power of communication lies in its ability to apply and interpret both its presence and the presence of communication originating from other actors. In the context of positive interventions, this applies to the form, tone, quality, and quantity of communication received and analyzed by a campaign from its target audiences. This process is commonly known as sentiment analysis – the process of detecting positive or negative sentiment in text. The focus of this effort is on the identification of certain types of 'polarity' that can be inferred from within the study of text (e.g., positive, negative, neutral).

Sentiment analysis as a tool, however, goes further than the relatively basic process of polarity detection into areas of analysis that interpret specific feelings and emotions (e.g., happy, sad, angry), urgency, and even intentions (e.g., interested versus not interested). These can be further branched into distinct processes that seek to grade different sentiments against certain desirable and undesirable criteria, engage in emotion detection analysis from rudimentary text, perform aspect-based analysis of sentiment driven by specific factors or conditions, and multilingual sentiment analysis that looks at variations of the same sentiment in different language contexts. Key benefits of sentiment analysis as an approach include its ability to sort data at scale, real-time analysis, and employing consistent criteria from which to analyze. However, there are methodological issues with how this approach would be used in prevention spaces, as there are challenges with inferring tone with written words, especially in massive data sets and across cultural boundaries. Polarity is also a challenge with mid-polar terms that rely on tone, sarcasm, culture, idioms, and context. Sentiment analysis is a useful tool for judging individual or community valence when there are drastic shifts in polarity, but more often than not, it is not reliable at scale.

It is important to note that sentiment and effectiveness are not synonymous, particularly when they relate to interventions in the realm of violent extremism. One of the most common validity issues with measures of intervention effectiveness is the assumption that when participants "feel" something, it is evidence that the intervention is effective. That is not necessarily true. Discrete emotion theory tells us that different sentiments have different action tendencies, meaning that when an individual experiences a particular emotion, she is motivated to act in a specific way. For example, if an individual experiences anger or disgust in response to an intervention, she is likely to discount or mentally dismiss the intervention. In this way, it is critical to understand (a) the emotions that intervention participants experience, and (b) their action tendencies in association with those emotions. The lesson here is that all emotional responses should not be considered equivalent and that making this assumption can lead to counterproductive intervention design.

Following from this, the measurement of discrete emotions (or sentiment) can be undertaken in one of two ways. First, thought-listing protocols can be very useful. In a thought-listing exercise, participants are simply asked to "free-write" about their perceptions of the intervention. Typically, participants are provided about 10 minutes to do so. There are no rules or restrictions as to what participants can write. Following this thought-listing exercise, two or three coders look at all the ideas written by participants and code them. Although coding schemes can be designed differently, at a minimum they should include two measures: (1) 'relevant'/'not relevant' and (2) 'positive'/ 'negative.' Ideas coded as 'relevant' and 'positive' can be considered affirmative sentimental responses to the intervention. A second method of measuring sentiment in relation to an intervention is a traditional survey. Surveys can be designed in a number of ways, but a recommendation is to embed sentiments of interest in a larger emotional scale.

Generally speaking, measures of beliefs and attitudes in response to intervention practices need to be tailored to the intervention so participants can be asked specifically about elements of the intervention. For example, if the purpose of the intervention is to reduce conspiratorial beliefs, belief scales should ask about those specific beliefs; if the purpose is rather to affect attitudes in relation to a specific minority group, then the attitude scale should be tailored to measure whether participants have positive or negative impressions of that group.

A theory relevant to measuring beliefs, attitudes, and sentiment is Reasoned Action Theory. The premise of this theory is that behaviors stem from beliefs about a behavior, norms surrounding the behavior, and one's capacity for engaging in the behavior. In other words, exposure to certain messages can affect beliefs about violence, which affects attitudes about violence, which affects intentions to engage in violence, which in turn affects violent activity.

Perhaps the most sought-after aspect of project measurement and evaluation is the need for analysis that considers diverse touch points or nodes of narrative engagement and sentiment. This is where more advanced forms of sentiment analysis through the presentation of visualizations come to the fore. Analysis can consider overall sentiment, sentiment over time, sentiment by rating, and sentiment by topic. By putting these touch points together or as a suite of measurement tools, different effects and sub-effects can be identified and correlated to glean if a certain narrative has met its desired conditions or objectives. This approach also serves as a very effective way to measure change in different contexts and conditions. Such a process has fast become the norm for online interventions that rely solely on the power of communication to create an effect. A key tool in sentiment analysis is the manner in which tech tools can be used to "opinion mine" in ways similar to how data mining has emerged as a standard practice for large data sets. The opinion mining process standardizes the

analysis element in order to measure "impact" or results, but is not without its drawbacks, some of which can affect the overall success criteria of a project/campaign or specific narratives.

The challenges of a sentiment analysis approach are diverse and carry significant risks that challenge both CVE and counter-disinformation efforts in similar ways. Subjectivity and tone are two particularly difficult issues to determine when analyzing complex data sets from audiences who may have very different ways in which they culturally or religiously express their views and may require the need to localize certain ideas. In terms of polarity identification, issues relating to contextual realities must also be considered. Factors such as the use and subsequent detection/analysis of irony, humor, and sarcasm are still areas that require more effort and expertise in capturing and interpreting. The way comparisons are made and their criteria can lead to controversy when it comes to reliable measurements of sentiment. The interpretation and understanding attributed to the use of emojis/icons is also an underdeveloped area of this field and requires further work.

Some of the most pressing issues involve how a sentiment analysis framework defines 'neutral,' which can be seen as a relatively subjective choice. This also links to the issue of human annotator accuracy versus machine learning systems using algorithms to both detect and interpret data. Despite these drawbacks, sentiment analysis has many potential avenues to follow with respect to the issues of interpretation and measurement of sentiment change, which are still key indicators of the possibility of behavior change. It is also noteworthy to point out that extremists in various guises have become more and more adept at harnessing sentiment analysis for their own narrative propagation needs. In the next section, a more nuanced discussion of these possibilities is introduced.

An area of the reach and impact debate that is full of rich potential is ensuring that measurement of sentiment change considers fully the distinctive nature of sentiment that is culturally specific, driven by more individualistic tones, and its origins and manifestations in ideological underpinnings. When these three nodes align within the expression of (for example) one person, the sentiment change analysis requires both flexibility and a degree of sophistication to interpret language, form, and tone in an equally nuanced way to its originator's levels. This can also be applied to sentiment that is more difficult to clearly categorize in terms of positive or negative, etc., due to it posing a set of complex assertions that may fall in between affirmative, rejectionist, challenging, or questioning tones of language/sentiment. This is particularly relevant to complex interventions with audiences who may be sympathetic to elements of extremist thinking (conservative or literalist thinking) and apathetic to other elements (worldview or violence). Such complexity brings into focus the issues related to subjectivity, the need for weight to be given to the contribution of indigenous studies in CVE efforts, and the relative differences between various groups in terms of power and influence or representation even.

A way through these challenges is to apply the study of semiotics alongside sentiment analysis for an even richer picture. Semiotics offers a study of signs and symbols that locate their meanings and values within specific normative and cultural contexts.[19] This enables certain language and expression motifs (colors, words, terms, icons, and symbols) to be attributed interpretive values based on the cultural conditions that create the value in the first place through shared experiences, stories, and communities. A classic example of this would be how the term 'conservative' holds many meanings in different cultural, religious, and political contexts, so it cannot be given an overriding interpretive meaning without risk. When combined, certain colors and symbols

19 "What is Semiotics?," Sign Salad, 2021, https://signsalad.com/our-thoughts/what-is-semiotics/.

create different meanings in different conditions (such as flags and slogans) and may serve very different purposes as a result. Both design and planning as well as evaluation and measurement stages can take advantage of semiotics to create a more in-depth awareness of possible outcomes and what these may mean. The more culturally competent the design, the more likelihood there is of impact and sentiment change being valid and able to direct future needs and interventions.

This all takes the issue of sentiment change and its analysis into a more comprehensive and defendable space and allows narratives to be more effective in the operating environment and not just in the online domain. The key to this process is seeing sentiment and change as phasic inputs created during the lifespan of a project's cycle. Capturing sentiment change is a process and sentiment analysis as well as semiotics are tools to aid this endeavor.

Section 2: Best Practices for Audience Targeting

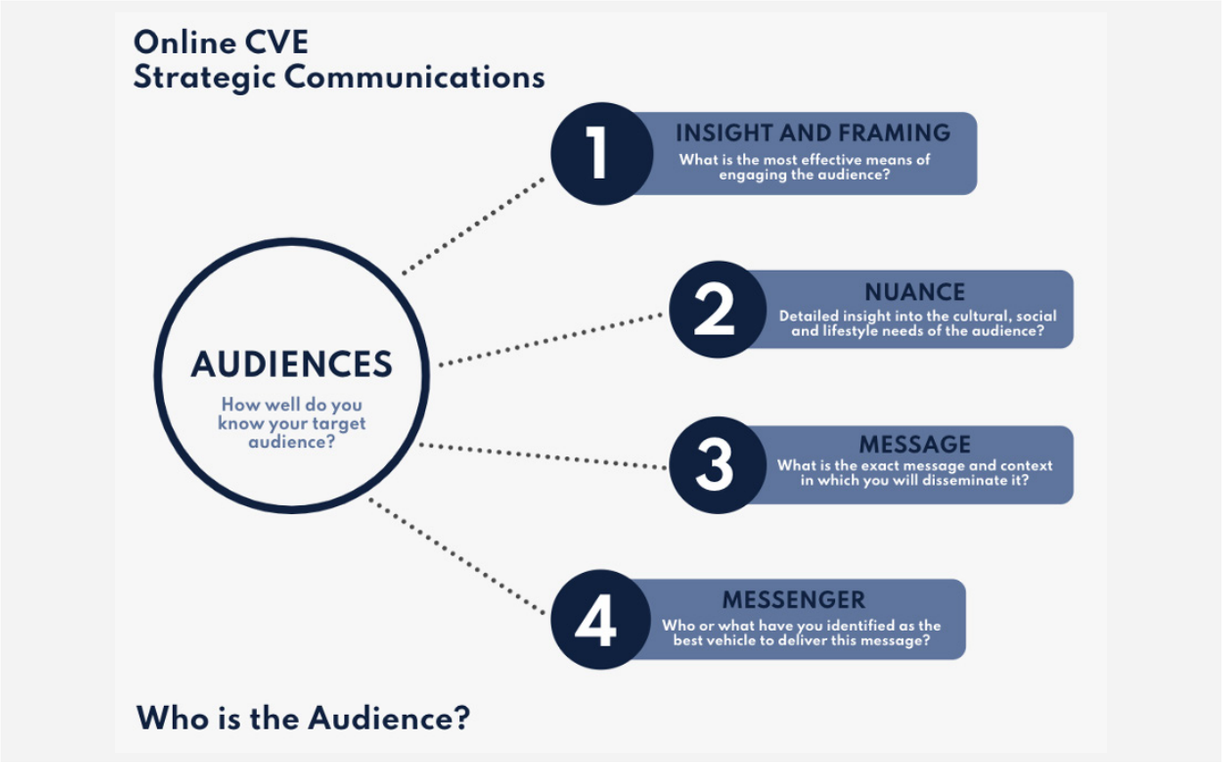
Audience targeting, both from a conceptual and practical perspective is an effort always in evolution and transition based on available tools, knowledge and access. However, certain general understandings of how best to approach do exist, in the shape of moral, ethical, transparency and sensitivity/safety needs. First and foremost, a human rights and rule of law lens should be at the core of any audience targeting. Such processes and understanding naturally vary, depending on who the actor is and the area of operations chosen for the positive intervention. From a CSO/NGO perspective, this process should be intertwined closely with the offline relationships/networks that are already in place or can be used to create new networks for specific project needs. In such contexts, the building of trust, meaningful engagement, and mutually beneficial relationships with offline communities allow the CSO sector to not only understand community needs but also translate these into tailored audience segments for the purposes of communication efforts.

From a planning perspective, audience targeting needs to retain a crucial position within related and specific aspects of project design. The basics of strategic communication theory began from a simple yet effective trident of "what, to whom, how" while nowadays there is added emphasis on "where." The question of "to whom" represents the all-important need for identifying, testing, and confirming the right audience has been found for the message to be disseminated to (and not always in a geo-location sense, hence the rise of "where"). With more and more technological tools at the disposal of many different actors in this space and the rise of the use of information as a weapon in influence and disinformation, it can be argued that audience targeting is now fast becoming the key to success for interventions based on communication.

The backdrop to this emergence can be sourced from the manner in which digital identity and lifestyle choices have given classical understandings of preference and choice a radically new sense of meaning and direction. Multiple-platform identities, online shopping and consumerism, online community networks, and the use of gamification techniques in different contexts have all helped to shape a more flexible understanding of what 'audience and taste' may mean. This increased flexibility offers the individual massive choice, but this equally creates additional layers of nuance for those entities seeking to attract, target and engage these audiences. Regrettably, QAnon has been particularly adept at using gamification to attract audiences, and this suggests innovation must continue to drive CVE development needs. In essence, the idea of 'an' audience is now something that needs various caveats attached to it to enable more robust and confident targeting, as merely assuming homogenous audiences still exist would be a mistake. Despite this increasing complexity, audience targeting still hinges on two foundational elements being conducted effectively: identification of a problem set

and a robust manner through which insight informs strategy. Audience targeting and insight require detailed processes of research, engagement, and segmentation in order to offer the messaging the best chance of engagement and impact.

Figure 4: Audience Targeting



For a positive intervention to have a chance at a relevant impact, the correct weight to the identification of the problem set must be added. Too often, problem sets become rudimentary and basic issues that are based on 'go-to' terms and industry understanding, rather than clear, robust, and defendable decisions regarding why a need for an intervention exists. In terms of mass audience campaigns, terms like "resilience building" or "increasing civic participation" have important rationales behind them, but when used in smaller and more tailored areas of operations, these terms represent different interpretive and assumptive meanings. In conflict zones or areas of instability, resilience building can mean being able to increase community support networks for greater self-reliance, whereas in other contexts it means the creation of networks or even building greater levels of critical thinking.

This applies equally to assumptions made about audiences within the academic field that may (for example) misconstrue vulnerability to extremism by suggesting environment or ideology as being a more relevant cause for fostering extremism or radicalization, whereas om actual real-world contexts environment and ideology are very difficult to separate as singular causal factors. The debate between Olivier Roy and Giles Keppel is a good example of this kind of discourse.[20] In choosing only environment or ideology to base both audience targeting and project objective setting, messaging, credibility, and impact can be adversely affected. 'Insight' and 'audience' are therefore not catch-all terms. Insight into what issues, on whom, and with which tools (e.g., tech, subject matter expertise, partnerships) involve scoping questions that require holistic treatment to be able

20 Adam Nossiter, "'That Ignoramus': 2 French Scholars of Radical Islam Turn Bitter Rivals," New York Times, July 12, 2016. https://www.nytimes.com/2016/07/13/world/europe/france-radical-islam.html.

to create the depth of information from which to put a strategy in place for a campaign.

Another useful way to approach audience targeting is to reverse the research parameters in order to formulate ideal and worst-case scenarios for the intervention life cycle. A small community-led intervention should consider broader, more macro audiences as ideal types in terms of desirable conditions, whereas a larger campaign with significant resources can consider creating clusters of micro-audiences who may have some key overlaps in terms of need and behavioral patterns that can help to better create and refine messaging through nuanced narrative and tone. This approach is a variation of the traditional use of A-B testing model but adapted to offer audience targeting a more robust and nuanced approach. With the rise of a more flexible and multi-layered form of online identity coming to the fore, granular level audience segmentation efforts and more nuanced audience insight processes enable projects to refine, identify, cross-reference, and target more relevant audiences through eliminating assumptions and relying on data driven feedback. This could be as simple as adding 'pattern of life' metrics into audience targeting, that capture lifestyle, network affiliation, tribal affiliation etc, in order delve deeper into possible entry points for various audiences who may share some similar factors, but are divergent on others.

As a result, being able to plan content and narrative themes can be processes developed with much greater levels of confidence when information mining lies at the heart of both conceptual and tactical considerations. In a traditional sense, demographics include age, gender, ethnicity, and location. These metrics offer scope to target certain audiences in certain places who may share similar features. If, however, lifestyle choices and network affiliation are considered, new layers and levels of nuance from 'similar' groups emerge that can radically alter both targeting and messaging needs. In the context of CVE efforts in the online domain, it is rarely considered that those deemed to be more 'vulnerable' to extremist messaging or ideology are also in today's world the same people who have a multitude of different identity facets that can be accessed as means to address alternative or counter-narratives.

This is where tech tools can be such powerful and transformative assets for campaigns. Examples would include insights that capture creative lifestyles, entertainment and leisure, and networking affiliations. When information mining reveals relevant and intriguing engagement possibilities, the issue of providing content that may trigger the possibility of a cognitive opening on issues linked to extremism or harmful ideas has more subtle entry points through which to build engagement, trust, and longevity. However, the management and use of information mining techniques and practices raise ethical and risk-based issues that should not be overlooked. Such approaches continue the trajectory of interventions having more nuance and adaptive thinking behind them in order to achieve outcomes that are behavior-change oriented. Essentially, in this way both the planning and delivery phases of a campaign can take a more considered approach to how knowledge, attitudes, and behaviors are initially understood and then shaped through content. Although the ways in which the commercial campaigns sector approaches these issues may differ slightly, the fact that they operate under the same assumptions suggests there are tips and tricks to be learned from the approaches of other sectors.[21]

Ethical Issues Around Targeting Vulnerable Audiences

The advent of the European Union's General Data Protection Regulation legislation has given greater layers

21 John Lee, "5 Practical Tips to Step Up Your Audience Targeting Game," Search Engine Journal, July 2, 2020, https://www.searchenginejournal.com/5-practical-tips-to-step-up-your-audience-targeting-game/373428/#close.

of safety to online users and given the commercial sector much to consider in terms of how audience insight and targeting can now be pursued from legal, moral and ethical standpoints. Historically, micro-targeting capabilities were able to use mobile/digital advertising criteria that allowed for highly specific forms of audience preference to be created via web-advertising and 'ad- displays'. For the commercial sector, this is now even more advanced via micro-targeting tools that use third-party data mining capabilities to be able to tie smartphone data (cell numbers and handsets), IP data, and geo-location tagging in order to offer highly targeted user profiles where required. Seen as standard practice in the commercial sector and the means by which the world's biggest corporations are able to attract customers they believe will be repeat, loyal and long-term prospects, this technology can easily transfer to the online intervention space for strategic communication needs. However, some ethical issues emerge in this context and questions of transparency, attribution, and intent permeate the technology versus engagement versus objectives debate.

This is of particular relevance when the issue of vulnerability among individuals and groups (and its many facets) is considered. One of the first factors that warrant attention is how the term vulnerability is used within the online intervention space, considering the improbability of any actor possessing a complete and holistic assessment of the level of vulnerability present in a group or person. Another consideration stems from how data mining, storage, handling, and aggregation issues coalesce with working in and around vulnerable groups/audiences. The need for robust quality assurance, security protocols, and risk management systems to be in place is without doubt a core need in this regard, alongside proportionality and transparency.

Vulnerability is an equally descriptive and relative term, and as such is often used to make grand assumptions about target audiences without a comprehensive understanding of how risk and protective factors affect people in similar socio-economic circumstances in different and divergent ways. In areas of conflict and suffering, this issue is still one that cannot be taken at "face value" when content and narrative are the only tools to achieve positive change. Project planning for online positive interventions needs to consider how to first mitigate against any known vulnerabilities through adequate risk management systems, potentially offering signposting and safeguarding support that is either directly or indirectly woven into content design and narratives as well as building tolerance within the project's scope to engage with any emerging risks or vulnerabilities that may arise once the content is live and being engaged with. Addressing vulnerable groups online is linked directly to real-world vulnerabilities that nevertheless manifest differently in each case. A project plan should include an impact assessment phase at both the initiation and end-stage, while also collecting midway point information on how content, audiences, and wider issues have collided or combined to either increase or reduce known vulnerability levels.[22]

Use of Influencers as Credible Messengers for Audiences

The issue of working (or seeking to work) with credible messengers has traditionally been based on access, levels of trust, and risk versus benefit considerations. The rise of the term 'influencer' has given the issue of credibility greater bandwidth, but not without some associated issues. The central one is that attempting to define an influencer is still very much an area housed within the commercial marketing sector, and variances of meaning regarding this term are numerous. A broad understanding of the term suggests an influencer is someone who is able to get relatively large numbers of followers to follow through with purchasing or

22 Jess McBeath, "Supporting Vulnerable Groups Online," UK Safer Internet Centre, 2022, https://saferinternet.org.uk/guide-and-resource/supporting-vulnerable-groups-online.

supporting a specific product, service, or cause.[23] Within the CVE context, credibility has morphed to mean different things, given that in specific areas of work deradicalization and hate speech could mean vastly different things. An example would be the use of former extremists ("formers") to help deradicalize individuals, as per the UK governments Channel program, versus using an NGO/CSO to communicate a message to counter hate speech by virtue of its community access translating to mean 'credibility.' It is also important to acknowledge the scarcity of openly available tools for measuring the specific impact an influencer or credible messenger may have had within a campaign. This issue needs addressing if the use and role of influencers is to become a data and research-driven practice within CVE for all actors within the field.

In an era where emerging social and fringe movements are hard to differentiate from each other, extremism and grievance have become more entrenched inside projectionism and "cancel culture" under the umbrella of the so-called "post-truth" era. This is perhaps best demonstrated when the terms activist, activism, and active participation are reviewed in the guise of online communication, engagement, and action. This applies to how differentiation between extreme views does or does not correlate to violent extremism and where the nexus points exist for policy and strategy considerations at governmental and commercial levels. If the rise of influencer culture is considered against these terms and contexts, it has to be taken as accepted that an influencer is given this moniker because there are identifiable and known metrics on this person's social network presence and reach. The term influence is closely allied to the term persuasion and the basis for much commercial, strategic, and ad-hoc communication is clearly to achieve some sort of persuasion-based impact, whether through subtle or overt influence regarding knowledge, attitude, or behaviors. If the lessons gleaned from audience targeting are taken into account, influencers are most influential when they are able to persuade others to make certain choices, think certain things and behave in certain desirable ways. Provocation strategies employed by digital influencers could therefore be closely studied and then adapted for the purposes of online CVE efforts.

What qualifies someone to be called an influencer in the CVE space is a subject that requires more research and understanding. This can range from being a former, an expert, academic, fashion expert, sports person, etc., and is of course entirely dependent on the context. In the CVE space, what isn't up for debate is the idea that a genuine influencer is someone who is able to exert some sway over the ideological impact, environmental conditions, and cognitive capacity of a stranger online who, regardless of having some link to the influencer, may have several other more relevant influencers and influences impacting their daily opinion-forming and action-taking. Influencers are worthy of consideration if they are demonstrably able to exert measurable influence over the choices of others, which can be traced and measured to confirm the causal relationship.

Depending on the objectives of the campaign, the level of influence is a key factor in looking and harnessing influencer support for a project. There are degrees and variations of influence at work in the CVE space, which means the choice of influencer is a crucial one, simply because a level of influence does not constitute credibility automatically in the complex environment of ideology, narrative, groupthink, and psychological motivations. The more nuances the era of digital identity has erected means influence and credibility are not immediately aligned features of any campaign seeking to undermine an extremist narrative or movement. The wrong influencer can create more vulnerability and take audiences further from safety, and the notion of the credible messenger has also become too much of a catch-all term when choice and preference have become so expansive, dynamic,

23 Jenn Chen, "What is Influencer Marketing: How to Develop Your Strategy," Sprout Social, September 17, 2020. https://sproutsocial.com/insights/influencer-marketing/.

and disposable. For example, a religious conservative influencer in practice may have been seen as a way into influencing vulnerable groups away from violent extremism but can equally be far removed from the values and norms of a certain society in regards to respect for rule of law, freedom of speech, and democratic processes. Another even more significant issue is how much influence fear and insecurity play in how audiences take in or ignore the overtures of "credible" messengers or influencers, where VEOs and insurgencies hold sway over daily life. From a community context outside of the CVE space, what 'credible' means for the role of the credible messenger[24] can be an interesting contrast to the CVE-specific perspective.[25]

This leads to another important facet of this issue, namely the assumptions behind the term credible. Actors can attribute the label of credibility to those who may have a certain persona online or reputation but lack the requisite coverage offline, and therefore can be seen as somewhat of a trojan horse to suspicious audiences. The different social worlds from where actors/campaign creators and audiences often come from mean that vanity metrics and persona play a bigger role in assuming a person's credibility than their network coverage and links to different audiences on complex topics. This can therefore create a "catch-22" situation once content is live and linked to a certain person/ influencer. The ideal scenario is one where the project team takes on the responsibility for a process of localized due diligence regarding a holistic assessment of the suitability of potential influencers and credible voices, taking into consideration both risk and benefit scenarios. Two areas within this issue that warrant closer attention are when youth audiences are involved and also the issue of working or enlisting "formers" to assist a campaign in a CVE context.

In a predominantly youth engagement context, two factors for consideration come immediately to the fore; what the term youth implies relative to age and how safeguarding and support (including peer support) are either incorporated or offered through third parties. In the online positive intervention context, the usual approach has been to have a lenient stance on age limits in regard to how a campaign or project determines its audience and participants for any workshops, focus groups, or peer-to-peer work. Youth can be defined in such contexts as anyone from 18 to 35 years of age. In many respects, this approach is logical given how the understanding of radicalization and vulnerability to it have fostered significant preventative efforts both online and offline. When looking at recruitment patterns of VEOs, this age range accounts for the vast majority of recruitment data available in the public sphere. It is also interesting to note that in other strategic communication arenas, where a specific cadre of VEOs or one main VEO is occupying territorial, social, and political influence, the same age range is known or referred to as military age males (MAMs).

What this contrast in terminology does is highlight that in most cases within the CVE sphere, youth are the key target audience, and therefore it becomes necessary to ensure that project parameters and pathways to engagement have fully understood and administered safeguarding and support inputs to protect against project, audience, and reputational risk, as well as impact assessments of content that is eventually present in the same environment as narratives of extremism, etc. This is where the "do no harm" principle comes into its own, by means of planning and programming that seeks to mitigate risk through identification, assessment, actions, and management as well as promote positive behaviors through socio-cultural norms and values. It is often possible to develop and support peer-led networks in the offline realm to facilitate such efforts and

24 "A Transforming Approach to Justice," Credible Messenger Justice Centre, n.d., https://cmjcenter.org/approach/.

25 Michael Jones, "Through the Looking Glass: Assessing the Evidence Base for P/CVE Communications," Royal United Services Institute, July 17, 2020, https://rusi.org/explore-our-research/publications/occasional-papers/through-looking-glass-assessing-evidence-base-pcve-communications.

over recent years this trend has also made its way into online interventions that access digitally connected user communities to teach them how to self-manage and develop narratives through a "network effect."

In the context of using influencers and credible voices in a youth-focused project or campaign, attention must be paid to the viral nature of social media influence, content, and attention/retention of interest. With the rise of platforms like TikTok, attention spans, optimum content style and length, remixed content, and repurposed content have all been moved forward into new common practices. This evolution has impacted the idea of influence, influencers, and reach in different ways. Where once positive interventions sought to retain audience engagement over longer periods of time, advances in technology and content creation now mean that bite-size content, both static and animated/ video, often has a three-to-ten second impression window and a thirty-to-sixty second duration at maximum. This gives rise to the proposition that the nature and tone of counter-narratives, alternative narratives, and strategic communication in general need to evolve in terms of their intended reach, impact, and desired end-state. Such a change would require a rethink in terms of storytelling, creating emotion and challenging/undermining harmful narratives in shorter and shorter spaces of time, when youth audiences in particular are of importance to a project's needs. In this respect, learning and adopting lessons from the commercial sector offer NGO/CSOs an effective way to leverage these new approaches to positive effect. This is why the more private-public partnerships that are based on equitable and transparent relationships, the more effective and innovation-driven CVE interventions will become. This applies to support and technical needs in equal measure, as matters such as content and its optimization are changing at a rapid pace.[26] The issue of how trends can shape and affect design and dissemination can equally not be understated or overlooked in this context.[27]

In regards to the use of formers as credible voices or messengers, much learning and outcomes (but not necessarily impact) driven information already exists. One2One, AVE Network, Abdullah-X, and Extreme Dialogue offer various important reference points for projects and organizations working in the CVE space. Some of the key factors in this context are that just by virtue of being a former, this does not necessarily mean the individual has knowledge, know-how, and tact in the realms of creativity, safeguarding needs, private-public partnership, and the role of tech. In many ways, this places even more responsibility on the project to create a conducive environment for the former to be able to "add value" to the narrative and for the audience to be able to resonate as well as relate to the presence of the former and their message. This again links to the increasing need for the private sector to assist with training, capacity building, and industry-standard tools to help formers become more accessible and properly utilized in order to leverage credibility, collaboration, and the continuation of initiatives.

Relatability stems from common themes or experiences, whereas resonance implies that a connection goes beyond its original purpose and influences other areas of interest or people of interest. For a former to be ultimately effective as a carrier of a positive message, tone, context, and attribution issues require careful consideration. This is because a brilliant counter-narrative can be self-sabotaged by it being attributed to parties and actors that cause suspicion within the audience, thus causing the former to lose face and the narrative to be attacked. The Abdullah-X project met with this exact fate as a result of enormous media attention, being

26 Claire Beveridge, "11 Easy Social Media Optimization Techniques to Start Now," Hootsuite, January 13, 2022, https://blog.hootsuite.com/social-media-optimization/.

27 "Social Trends 2022," Hootsuite, 2022, https://www.hootsuite.com/research/social-trends.

paraded around the global CVE circuit and eventually offering jihadist sympathizers the chance to frame it is a Western attempt to "undermine Islam." It became relatively easy to track the projects' media presence and tie it to organizations and individuals of interest for its adversaries. Language and dialect, cultural nuance, and visual presentation all play a role in how credible a message is and how credibly a messenger is received, even if this happens to be a former. With rapid changes in engagement and reach related issues, it may be useful for positive interventions to be planned with the idea of credibility only being able to be understood once the audience has decided how credible a person or message really is to them and for how long. The credibility issue has several other considerations within sectors such as capacity building, advocacy, and offering support services.[28] In regards to accessing and working with victims of terrorism, other considerations and sensitivities[29] come into play and require both management and tact to be able to activate and manage such relationships.[30]

Creating a Trustworthy Online Presence

When engaging in digital work, especially when the aim is to establish direct contact between an at-risk or radicalized person and a counselor/P/CVE professional, questions of credibility and trust building are closely linked to transparency. This means any professional wanting to engage in this field needs to invest time and thought into creating a convincing online profile that allows insights into their personality and background without including private information to maintain safety. A well thought through profile may be able to create a sense of intimacy and trust that can be able to bridge the distance even in digital settings. Professionals also need to be truthful in their communication about their professional affiliation and objectives. According to many practitioner experts, using anonymous profiles and false pretexts to start conversations with hard-to-reach target groups has proven to be unreliable and unsuccessful. Danish expert Christian Mogensen from the Center for Digital Youth Work has previously spoken about his visible track-record of publications on positive aspects of gaming and online cultures, which, according to him, increased the initial trust of "Incel" community members and enabled him to engage them. While this level of online exposure may not be possible/wanted for regular practitioners, the organization or program with which practitioners are affiliated needs to ensure that their communication around their work supports trust building prior to engaging with a project. The strategy on how a program or practitioner presents themselves needs to be adapted based on the specific topic area/phenomenon they are working on.

Facilitating Cognitive Openings

When looking at the core aims of any strategic communication effort, it goes without saying that some form of behavior change comes high on the list of desired end-states or goals. The shape and form of this behavior change is a different matter, and for this reason, when considering the need to impact audiences at a behavioral level, some form of knowledge and attitude building or shaping needs to occur in order for the behavior change process to be monitored and captured. This process comes together under the guise of creating a cognitive opening. In contrast to positive interventions seeking to create some form of cognitive opening in the opposite direction, extremist narratives can create ruptures in the cognitive alignment of

28 Marina Tapley and Gordon Clubb, "The Role of Formers in Countering Violent Extremism," ICCT, April 12, 2019, https://icct.nl/publication/the-role-of-formers-in-countering-violent-extremism/

29 "Counterspeech: Extreme Lives," Facebook, n.d., https://counterspeech.fb.com/en/initiatives/extreme-lives/.

30 Giullaume Deniox de Saint Marc et al., "Handbook: Voices of victims of terrorism," Radicalization Awareness Network, May, 2016, https://ec.europa.eu/home-affairs/system/files/2020-09/ran_vvt_handbook_may_2016_en.pdf.

individuals, resulting in cognitive dissonance (when existing beliefs and newly formed extremist beliefs clash and create inconsistency between belief and overt behavior or action). A cognitive opening seeks to redress inconsistencies and create a new window of thought, conceiving different scenarios and applying reason to emotions and narrative cues. Within the positive intervention context, there are numerous ways to seek this type of opening and various tactics that align themselves with innovation and creativity to achieve the desired outcome, which is to take audiences further away from harmful ideas and actions. In section 3 of this report, further exploration on the various ways and means of seeking a cognitive opening are discussed in order to build upon the conceptual and practical implications of this type of work.[31]

Incentivization

A relatively obvious means to seek a different cognitive path for an individual or group is to offer some form of incentive or reward for certain types of behavior or actions. This can be through increased civic engagement at the local level, competitions, prizes, innovation labs and tech-led challenges, or tangible routes to manifest aspiration levels. The general aim of such an approach is to provide a stimulus of motivation, confidence, and self-reliance that ideally can combine to create a greater sense of personal agency. An incentive-based approach can be used in many of the standard positive intervention objectives already mentioned (e.g., deter, intervene, prevent, build and empower) and is perhaps most effective when resilience building and capacity building aims are at the forefront of the project/campaign's desired outcomes.[32] Another example of a delivery model in this context is to use peer-to-peer efforts, which can unite engagement, empowerment, and localized efforts.[33] PeaceGeeks' project with the Meshkat community in Jordan also offers a highly localized and community-focused approach to tackling hard-to-reach groups and addressing sensitive topics.[34]

Faith as a Facilitator of Cognitive Openings

Using ideologically-based narratives for introducing influence via non-traditional or non-institutional religious beliefs and ideas is a more complex process in the positive intervention space given issues with representation and matters of authority and authenticity alongside accuracy of messaging in regards to meaning and interpretation. Having said this, much of the apprehension around ideologically-based messaging or content stems from a lack of contextual knowledge as to how extremist narratives misuse faith for sinister means and ends. It is very difficult to challenge harmful narratives that include distorted renderings of religious text if the response is not created in a clear, credible, and meaningful context and in a compelling and creative manner. There have been several effective attempts to harness ideologically-based narratives that have accessed figures of religious authority, historical narratives, creativity, rhetoric, satire, and critical thinking tools to deliver predominantly religious messaging aimed at undermining extremist ideology and narrative. Such examples tend to suggest that elements of religious understanding can be packaged in various narrative forms to encourage self-reflection and questioning, and as a result some can contribute to positive cognitive openings.

31 Casal Bertoa and Jose Rama, "Polarization: What Do We Know and What Can We Do About It?," Frontiers, June 30, 2021, https://www.frontiersin.org/articles/10.3389/fpos.2021.687695/full.

32 "Counterspeech: Resiliency Initiative," Facebook, n.d., https://counterspeech.fb.com/en/initiatives/resiliency-initiative/.

33 "Counterspeech: Peer-2-Peer," Facebook, n.d., https://counterspeech.fb.com/en/initiatives/p2p-facebook-global/.

34 "Meshkat Community Amplifies Arab Voices for Social Inclusion in response to online hate, violence, and polarization," PeaceGeeks, n.d., https://peacegeeks.org/meshkat-community.

Two important yet little-known examples of when faith-based narratives opened avenues for the non-violent rendering of certain textual sources are the Mardin declaration of 2010 which addressed and refuted disinformation efforts with some significant faith-based authorities.[35] Similarly, the Amman message convened two hundred Islamic scholars from around the world with the aim of differentiating mainstream Islamic principles from extremist renderings of the faith on core subject areas that are often exploited and intentionally misused.[36]

However, the issue remains that even such important examples as those noted above have not been repacked and repurposed in creative online ways to have a bigger impact. This perhaps speaks of a lack of awareness and skill set in some areas of CVE work to adequately understand and then juxtapose the nexus between faith, motivation, and intent in many extremist ideas and narratives as well the challenge of being able to bring nuance to harmful viral ideas. This issue is even more complicated when the online context is the terrain in question. Issues relating to trust building and specific skill sets linked to digital literacy and practitioners' understanding of these (e.g., informed choice making, source verification, online safety, etc.), are important issues for the CVE community as a whole to continue to develop in the hope of bridging gaps. More work is needed to bring better faith-based understanding and mainstream narratives from within faith contexts into online CVE planning and delivery. Lack of specialized religious or cultural understanding can no longer be seen as an excuse for poorly designed ideas that can have adverse effects on faith-based communities. A role for the research and tech community could emerge as a keeper of repositories of up-to-date learning and resources from such contexts in partnership with viable CSOs.

Lack of Digital and Social Media Literacy Among "Traditional" Practitioners

It remains a major challenge that a significant portion of highly experienced practitioners, especially from the secondary and tertiary prevention levels, are themselves not experienced online like their digital native target groups. Tech companies could invest in training practitioners to improve their capabilities within the digital world and support P/CVE efforts at all levels. Additionally, practitioners require insights into the latest developments of extremist narratives and activities online so they can act and react in a timely manner. And even beyond that, they need to be aware of the key topics and trends that currently move their target groups. There is a lack of concise practice-oriented analysis regarding such developments that practitioners can refer to during their daily work – ideally with a country- or even region-specific component. Most practitioner organizations do not have the resources to do this themselves. A good practice in this area is the continuous basis monitoring implemented by Modus|Zad in Germany (funded by the German Federal Agency for Civic Education).[37] The project produces short monthly/quarterly practice-oriented reports based on a previous analysis of relevant extremist actors online. As of 2022, additional dissemination tools (such as webinars and workshops) are being implemented.

Use of Humor

35 Shaykh Abd al-Wahhab al-Turayri, "The Mardin Conference – Understanding Ibn Taymiyyah's Fatwa," Muslim Matters, June 29, 2010, https://muslimmatters.org/2010/06/29/the-mardin-conference---a-detailed-account/.

36 "The Amman Message," n.d., https://ammanmessage.com.

37 "Basis Monitoring (2021/22)," Modus | Zad, n.d., https://modus-zad.de/schwerpunkte/monitorings-trendanalysen/basis-monitoring-2021-22/.

Humor can be a very effective tool for engaging audiences, building momentum in user following and reach, and communicating complex issues in accessible and memorable ways. Its usage and efficacy depend largely on the ability of content creators to form and maintain "touchpoints" of interest and then curiosity through a medium mainly used for entertainment. Humor that is delivered through comedic and satirical mediums does need appropriate sensitivity checks to mitigate against negative reactions and the risk that messaging is taken out of context and reused by extremists or harmful actors to attack the very auspices of counter-extremism efforts.

Use of memes has proven to work well with some target groups. They provide a low-threshold opportunity to touch upon difficult topics and to elicit reactions and a debate from the target group, which can lead to further engagement and sometimes even direct contact and counseling. When thematizing difficult and potentially disruptive topics which are bound to raise emotions among the target group in a non-humorous way online, reactions often include considerable adversity, including hate messages and threats. Generally, however, whenever working with humorous or disruptive content, practitioners need to plan for time to observe and engage in the debates that arise as a result. Careful planning is definitely necessary regarding human resources, capability and capacity. There are examples of the use of comedy and humor within P/CVE adjacent fields.[38]

There is clearly scope for using comedy, satire, and humor within online and offline CVE efforts,[39] but it must also be acknowledged that such tactics carry risks, which need to be managed, mitigated, and accepted with a degree of caution and consciousness.[40]

Crisis Points and Deployable Content

Crisis communications require speed and often pre-prepared messaging to be combined with context specific needs. The link between this need and audience targeting issues is based around the nature of the crisis and who the first-hand affected parties are, the larger community around them and then the national/ international context. Based on the emergency, defining who to communicate to is to some extent driven by who is most affected and then working backwards in order as suggested. In a moment of crisis, communication becomes a key interlocutor for both government agencies, communities, and mainstream media. The need for timely, accurate, and trusted accessible information becomes the primary objective in such situations. For those involved in managing and disseminating crises communications, other variables of equal importance are also part of the strategic framework that forms effective "crisis comms." A key need is to be reactive in a manner that is appropriate to the actor's position within the overall context. For example, a government agency may be required to provide timely and accurate information relating to emergency response needs, emergency numbers for people to call, information as to what citizens should and should not do, where to go and where to avoid going, etc. The tone and choice of language and messaging style needs to be informative and avoid undertones

38 Shezo Media, Facebook, n.d., https://www.facebook.com/ShezoMedia/videos/; Fata Byyano, Facebook, n.d., https://www.facebook.com/Fatabyyano; PRODUCTIONS ON OUR PALETTE, Kharabeesh, n.d., https://www.kharabeesh.com/entertainment.php.

39 Inari Sakki and Jari Martikainen. "Mobilizing Collective Hatred Through Humor: Affective–discursive production and reception of populist rhetoric," British Journal of Social Psychology 60, no. 2 (April 2021): 610–634, https://bpspsychub.onlinelibrary.wiley.com/doi/full/10.1111/bjso.12419; Olga Khazan, "The Dark Psychology of Being a Good Comedian," The Atlantic, February 27, 2014, https://www.theatlantic.com/health/archive/2014/02/the-dark-psychology-of-being-a-good-comedian/284104/.

40 Tom Dreisbach, "How Extremists Weaponize Irony to Spread Hate," NPR, April 26, 2021, https://www.npr.org/2021/04/26/990274685/how-extremists-weaponize-irony-to-spread-hate?t=1651651473874&t=1652439765302.

that could exacerbate already tense issues further. It is important to acknowledge if information is simply not known.

The response to the 2019 terror attack in Christchurch, New Zealand remains both a best practice example for crisis communications, but also an example of how to communicate information without compromising the integrity of the office of government, the needs of the grieving and concerned families, and being steadfast in naming what the act is and isn't in order to get ahead in the war of narrative and information race that sadly commences in such instances. This example perhaps best demonstrates the value of quick, clear, and robust communication needs in crises situations, and how this can galvanize the community in a positive direction. A key example how addressing engagement and communication in points of crisis, where sentiment and reaction are very sensitive issues for all involved, was the "Je Suis Charlie" effort in the aftermath of the terrorist attack in France.[41] In the immediate days and subsequent outpouring of emotion and need for healing and grief, coverage of the campaign was very much supported and seen as a necessary statement of unity and strength in a time of loss and anger.[42] Similarly, in Nigeria the "Bring Back Our Girls" campaign achieved significant global coverage and ensured the actions of Boko Haram were contextualized from the victims' perspective rather than fuel more coverage for the group's aims.[43] The transference of sentiment and support made this campaign an effective use of social media and network effects at a time when confusion, anger, and fear had gripped the community affected.

Other actors (Government agencies) may have roles that require their communication to be of a reassuring nature, or the need to reaffirm certain core values to worried citizens. Messaging that sends out a clear affirmation of a willingness to challenge, undermine and defeat violent extremism may well be the domain of specialist NGO's and CVE practitioners in partnership with Government and or tech sector actors . In such instances, having stock content, archival content, and rapid response creative treatments available at short notice saves both time and resources for already overstretched community partners and stakeholders. Content that can be repurposed and repackaged may stand a better chance of winning the narrative war that ensues in the aftermath of such incidents. The different stages and needs of responding to crisis points also require consideration and awareness in order to ensure consistency and efficacy are maintained.

What Do We Want Our Target Audience to Do Once the Campaign Is Over?

As mentioned previously, a primary aim of community-centric strategic communication is to be able to influence audiences based on a desired end-state or goal. In the CVE context, this focus can take the shape of increased resilience building, rejection of extremism, or greater levels of critical thinking to name a few. When online interventions are allied to relevant offline interventions, the question of longer-term impact or legacy takes on additional significance. Taking the example of having adequate support or redirecting the energy of participants for offline interventions, this allows audiences to seek and access support services or community-linked pathways to social, employment, and cultural inputs that can build a more sustainable impact through

41 Jonathan Ervine, "Five Years on from the Charlie Hebdo Attack," The Conversation., January 6, 2020, https://theconversation.com/five-years-on-from-the-charlie-hebdo-attack-je-suis-charlie-rings-hollow-129151.

42 Mukul Devichand, "How the World was Changed by the Slogan Je Suis Charlie," BBC News, January 3, 2016, https://www.bbc.co.uk/news/blogs-trending-35108339.

43 Joe Parkinson and Drew Hinshaw, "How the 'Bring Back Our Girls' Tweets Changed a War in Nigeria," Wall Street Journal, February 20, 2021, https://www.wsj.com/articles/how-the-bring-back-our-girls-tweets-changed-a-war-in-nigeria-11613797261.

longer-term engagement. Within this model of intervention, a key element of the relative success of such efforts is how projects have infused objectives within a community-oriented environment through relevant actors and organizations who are able to engage and retain more "hard-to-reach" groups. Before this model is discussed in the context of the online space, some elucidation of the premise of positive interventions and their proposed outcomes is needed.

A key factor in this discussion is based on the scope, limitations, and bandwidth of communication as a tool to reduce the allure of extremism, harmful socio-political movements, and violence as a means to seek change. While it is acknowledged that radicalization and deradicalization are still very much contested concepts, both serve as useful narratives in their own right to explore both the potential and pitfalls of solely communication-focused interventions. The reason for this suitability is that many CVE communications efforts are based at the conceptual and design level on the propositions advanced by radicalization studies. In order to therefore delve into the longer-term ambitions of such efforts, in terms of impact on audiences and conceptual and real-world variances, anomalies and assumptions need to be identified.

A simple example of the over-reliance on radicalization discourse can be understood when one considers that the concept of socialization is barely mentioned in the literature in regard to possible processes by which to safeguard vulnerable communities from extremism.[44] Equally, there is an enormous body of literature available that links the allure of extremism, religious cults, violent social movements, and insurrectionist networks to broader themes around the need for identity, a sense of belonging, and a sense of purpose, loyalty, or duty among many who regrettably join such entities. This suggests that if radicalization is seeking to explain how this process occurs, no standard blueprint can be applied as no two journeys into extremism can ever be completely identical. What happens prior to the onset of radicalization is rarely considered, yet in many ways this pre-extremism phase is exactly where many communications efforts in CVE are often basing their objectives (albeit indirectly), as will be explained below.

A UNDP-commissioned study on drivers, incentives, and the tipping point for recruitment by violent extremists in Africa found several common "push" and "pull" factors around locality, opportunity, and forced/exploited incentives at the heart of VEO recruitment.[45] Such findings are important and need to be understood in relation to the role of social/religious cults and psychology/identity needs also.[46] The use of identity and belonging narratives need to be better understood in order to inoculate communities against extremist renderings through redirect methods and more open discussion of these themes of these issues.[47]

Whenever a communication-focused intervention is seeking to reduce support for extremist narratives, reduce extremist recruitment and provide alternatives, there are often narratives designed and disseminated that seek to increase knowledge, shape attitudes, and ultimately change behavior among audiences. The main aim is

44 Paul Hedges, "Radicalization: Examining a Concept, its Use and Abuse," International Centre for Political Violence and Terrorism Research 9, no. 10 (October 10, 2017): 12–18, https://www.jstor.org/stable/26351560?seq=1.

45 "Journey to Extremism in Africa: Drivers, Incentives and the Tipping Point for Recruitment," United Nations Development Programme, 2017, https://journey-to-extremism.undp.org.

46 Daniel Verana, "A Look Behind the Curtain of Cult Psychology," The State News, October 29, 2020, https://statenews.com/article/2020/10/a-look-behind-the-curtain-of-cult-psychology.

47 "Identity, Belonging and Extremism," European Union: Organising Intercultural and Interreligious Activities: A Toolkit For Local Authorities, n.d., http://www.congress-intercultural.eu/en/initiative/215-identity--belonging-and-extremism.html.

to get audiences to act differently enough to be able to measure some form of reduction in extremist activity, presence, or support. If it is considered that all forms of extremism take root in and around "mainstream" social realities and many adherents of extremist thinking have already made an active choice (voluntarily or otherwise) to reject or move away from such norms and values, then attempts to dissuade them are offering the same set of norms and values that have been shunned already. In other words, when positive interventions or strategic communication efforts use terminology such as increasing critical thinking, rejecting extremist narratives, and becoming more aware or better active citizens in a strategic objective setting context, they are making a set of assumptions that a narrative on its own can impact and replace the allure of identity, belonging, and sense of duty that extremist ideology offers. This point has two important real-world implications for interventions in today's crowded information space: measurability and specificity.

Extremism offers not only identity, belonging, and sense of duty narratives, but in many instances can offer these through networks, roles, and opportunities for self-gratification. This "package" ultimately replaces traditional notions of family, friendship, and socio-economic stability that form the basis of modern liberal democratic societies, which is in and of itself an issue given that such societies are not the places where the most urgent interventions are needed. Yet for reasons that are beyond the scope of this report to explore in any great detail, this underlying contrast of worldviews still affects how effective positive interventions can be. This point of divergence in understanding and context is then amplified still further when it is considered that the presence of extremist thinking has permeated several mainstream narratives and latched itself on to the rise of disinformation and an increase in harmful actors who are difficult to categorize either as extremists or as activists. Put another way, it has become difficult to target vulnerable groups when mainstream society is now the most vulnerable group and geography plays less of a factor in where needs are most pressing intervention-wise. The loaded nature of the way in which the term extremism is used is in and of itself problematic when seeking to define parameters of interventions, apply legal context, and adhere to human rights.

The rise of technology and the role it plays in offering tools and means by which to proliferate extremist ideas and disinformation are issues that represent another challenge for civil society actors and the tech sector alike, alongside trying to challenge and undermine extremism globally. Conflict zones, areas where autocratic regimes operate, etc., are no longer the sole target for extremists. Between 2016 and 2022 alone, the landscape of influence, partisanship, polarization, and binary narratives has coincided with a sharp rise in new social movements, othering narratives, and cancel culture in many well-developed liberal democracies. How effective can a communications campaign be in this context when other actors and entities offer not just narratives but lifestyles and active participation?

This creates a set of 'relative uncertainties' about some of the intended aims of positive interventions that have already been mentioned in this section. How critically do we want audiences to be able to think? What if they think so critically that they oppose both the narrative being proposed to them and furthermore take steps to voice this opposition? What does resilience look like to someone already being offered a job and a sense of unity and loyalty from a VEO when the social reality they live clearly cannot satiate these needs through "hope" alone? How will having a greater sense of awareness of things negate extremism when levels of trust in authority and questions around legitimacy are now commonplace in state versus citizen narratives? What does citizenship look like in the absence of a clear and identifiable sense of security? What values systems does one draw upon in considering participation in a VEO?

Such questions raise again the need for positive interventions to be immersed in the lived reality of their

intended audiences. Such interventions should rely on highly developed levels of insight and expertise regarding the drivers and factors creating the problem set and able to pursue a suite of nuanced, timely, and compelling narratives that take advantage of the tools on offer within the tech sector (through improved private-public partnerships) in order to offer choice, a sense of digital "sanctity," and a policy of being present and "there" for audiences in their time of need. Interventions should look to legacy impact as much as initial impact, with as much future-proofing for content and access as resources and scale allow.

The quest for more private-public partnerships within the CVE field again goes to the heart of this debate and is something critical to the long-term success of efforts to counter extremism at the source.[48] How the role and remit of the CSO sector develop in CVE is seemingly tied to the ways the partnership grows over time to meet the evolving threats, risks, and needs of communities.[49] YouTube's Creators For Change program is a good example of how to bring tech and communities closer together in this regard.[50] Issues linked to the need for trust and credibility are of course paramount to this type of partnership and cannot be overlooked for 'quick wins' or vanity level metrics. [51]

Section 3: Turning Passive Counter-narratives into Active Strategic Communications

As previously discussed, increased critical thinking as an objective still remains popular within positive intervention design and content threads. The reason for this is based on the proposition that the higher the level of emotional intelligence and ability to think critically about information, choice, and outcomes will offer audiences a safety net from narratives that possess a more binary "them and us" focus. The PeaceGeeks initiative involved critical thinking through elements that include looking, assessing, and responding, then complemented it with Social and Emotional Learning (SEL) to work on emotional intelligence. The rationale is that critical thinking on its own is not enough to foster emotional intelligence.

The generic critical thinking approach champions the role of informed choice making and validating sources as a means to avoid falling into the echo-chamber trap and confirmation bias patterns so prevalent in extremist narrative. In order to make critical thinking a powerful tool in this regard, static content and narrative with a one-way engagement focus face an uphill task to permeate both curiosity levels and cognitive avenues to behavior change, simply due to one important component being missing: active two-way engagement. However, this is not the two-way engagement of classical approaches to strategic communication; this is a term used to describe a more immersive, process driven, and decision-making oriented set of exchanges. The main aim of this approach would not solely be the exchange or engagement, but also the experience (ideally a shared experience wherever possible). Ultimately, the human need for shared meaning and experience drives both extremism's allure (at least initially in many cases) but is also the key to challenging harmful narratives through

48 Alejandro Beutel and Peter Weinberger, "Public-Private Partnerships to Counter Violent Extremism: Field Principles for Action," Final Report to the U.S. Department of State (College Park, MD: START, 2016), https://www.start.umd.edu/pubs/START_STate_PublicPrivatePartnershipstoCounterViolentExtremismFieldPrinciples-forAction_June2016.pdf.

49 "The Role of Civil Society in Preventing and Countering Violent Extremism and Radicalization that Lead to Terrorism," Organization for Security and Co-operation in Europe, August, 2018, https://www.osce.org/files/f/documents/2/2/400241_1.pdf.

50 "YouTube Creators for Change," YouTube, September 19, 2016, https://www.youtube.com/channel/UCYJJpu7FLQqu788cusj6nlg/about.

51 Kurt Braddock and John Morrison, "Cultivating Trust and Perceptions of Source Credibility in Online Counternarratives Intended to Reduce Support for Terrorism," Studies in Conflict and Terrorism 43, no. 6 (2020): 468–492, https://www.tandfonline.com/doi/abs/10.1080/1057610X.2018.1452728.

audiences having the chance to immerse themselves in alternative ways of thinking as opposed to just being told about them.

Friction As Counter-speech

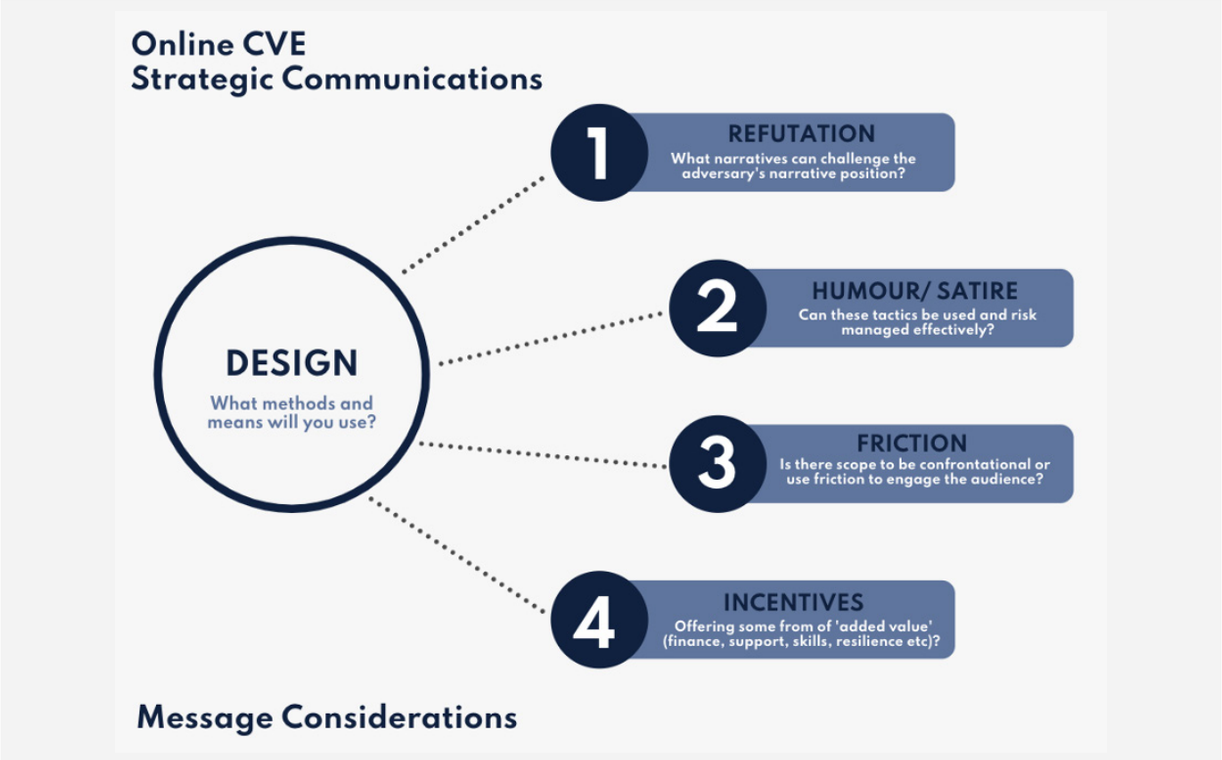
Within the positive interventions space, friction can be used in many ways for different means and end goals (e.g., undermining, refuting, questioning). One aim of using friction in this context is to create uncertainty and doubt about the authority and superiority of the extremist position through direct challenges to the auspices under which such claims are made. This aim is driven by the presence of counter-narratives that question, attack, and malign the assumptions, distortions, and misinterpretations that extremists use, turning these against them through intellectual, rational, yet relatable narrative content that causes "breakpoints" between idea and action in the target audience. The aim is to posit some ideation, but this alone cannot create a total rejection of an entire existing position that may not have used "logic" or "reason" to embed itself originally.

In "Five Considerations for a Muslim on Syria" by Abdullah-X, there are specific trigger points (e.g., questioning of motives, managing emotion, feeding curiosity) used in the narrative that contains elements of friction designed to elicit internal questioning of the justifications presented by ISIS as well as broadening out the conflict with issues related to a desire for power and control.[52] The goal of such a narrative is to give the user an opportunity to question his/her motives for considering harmful actions or ideas alongside being given intellectual tools to be able to better understand the "bigger picture." According to Kurt Braddock, "Messages like the one communicated in this video provide viewers with a source of pride that (a) they can identify with and (b) challenge terrorist ideologies and actions... Abdullah-X, who describes himself as having the 'mind of a scholar' and the 'heart of a warrior', shows young Muslims targeted by ISIS propaganda that they can be proud of their Muslim heritage, but need not affiliate with ISIS."[53]

Various tactics within this context can be used in isolation or as a set of content within the same project/ campaign context, as demonstrated by the figure 5 below:

52 Abdullah-X, "Five Considerations for a Muslim on Syria," Youtube.com, March 7, 2014, Ibid, https://youtu.be/tKKbydB4scA

53 Kurt Braddock, Weaponized Words: The Strategic Role of Persuasion in Violent Radicalization and Counter-Radicalization (Cambridge, UK: Cambridge University Press, 2020), 196.



Another use of friction in counter-speech is to "call out" extremists by isolating their tactics and actions as those of people who clearly do not value human life and the values that let societies flourish as opposed to perish. This works best when the positive examples are from within the same social, cultural, historical, religious, or geographical context. Such tactics can utilize "discrete emotion theory" to communicate messages that offer alternative yet compelling pathways to extremism through positive goal targeting, successful actions, and highlighting similarities between these approaches and the target audiences' core values. These examples suggest that friction can have subtle variances in how it is used, applied, and interpreted. One way to identify this tactic is by labeling such messaging as "positive friction," an example of which can be found in the "Syria Street Stories" campaign.[54]

Positive friction is perhaps best illustrated in the commercial advertising world, where the constant struggle between sportswear giants Nike and Adidas for market dominance uses differentiation and lifestyle narratives to offer "unique" choices to often similar customer bases.[55] Companies "challenge" their rivals through unique storytelling and values-based narratives, all the while undermining the same things in their rivals by identifying areas of divergence in image and lifestyle, etc. Nike has long maintained that "just do it" is about being both in the moment and an unapologetic mindset, in order to harness, access, and develop one's capability levels. Adidas, which also targets a similar customer profile, like to remind these potential customers about their brand's longevity, street cred, and cutting-edge persona (in terms of being both 'old-school' and innovative), using influencers and familiar faces to build brand value. They have a well-known strap-line "the brand with the three stripes," despite their actual logo having changed from the classic logo the strap-line is referring to. Both brands have strap-lines that mean different things, but still speak of "values" that most likely mean similar things in both the lifestyle and sports contexts. The friction element is hidden inside sophisticated

54 "Abbas's Story," YouTube.com, February 26, 2016, https://www.youtube.com/watch?v=4WeUQy972Kw&list=PLb3EHFgagz0xiR2tRNg-qahBdEHhW49-s.

55 Sofia Deneke, "Nike vs. Adidas: Whose Marketing Strategy Reigns Supreme?," Trig, July 27, 2017, https://www.trig.com/tangents/nike-vs-adidas.

and often indirect references to each other's brand without directly naming the other party.

What Is an Appropriate Amount of Friction?

This relates to both the appetite for the medium and the propensity to manage and mitigate risk. Direct confrontation through friction is technically designed to elicit a response, and this means that actors need to gauge and model the potential impact of friction on reputation, safety, and audience wellbeing levels. The rationale at use here is that friction begets more friction and this means having a willingness to be reactive, flexible, and attentive to changing dynamics and subtle cues. This is a scenario that must include comprehensive end-state planning and what decisive conditions need to be met in order for the use of friction to be considered a successful tactic and not a "flop," risk, or liability.

Therefore the use of friction carries both specific potential benefits while also significant risks. From a management perspective, friction runs the risk of snowballing and taking on a life and narrative angle of its own. If this includes the target audience becoming embroiled in a war of narratives and attacks on the other party, online safety issues quickly translate into offline risk of harm issues, given online identity is not always as hidden from prying eyes as people like to think. Project administrators in this case may need to have in place support structures and contact information for audiences to use in case of threats and trolling/harassment, as well as encouraging more digitally "savvy" account names and profiles, etc.

Managing and Measuring Impact of Friction on Audiences

In terms of measurement and impact, A-B testing with and without elements of friction in the content on the same audience offers an empirical way to demonstrate both the reach, engagement, and second level or sub-effects of the friction being employed. In some more sensitive operating environments, multiple channels carrying subtle variations in messaging and tone can also create a network effect for friction-based content that suggests support for the antagonist's point of view is smaller than support for the protagonist's perspective. Pooling the bandwidth of such content in terms of data mining for where the friction-based content appears in different online places also allows for reach and impact to be measured more widely than just on the official channels.

Building User Resilience in the Online Space

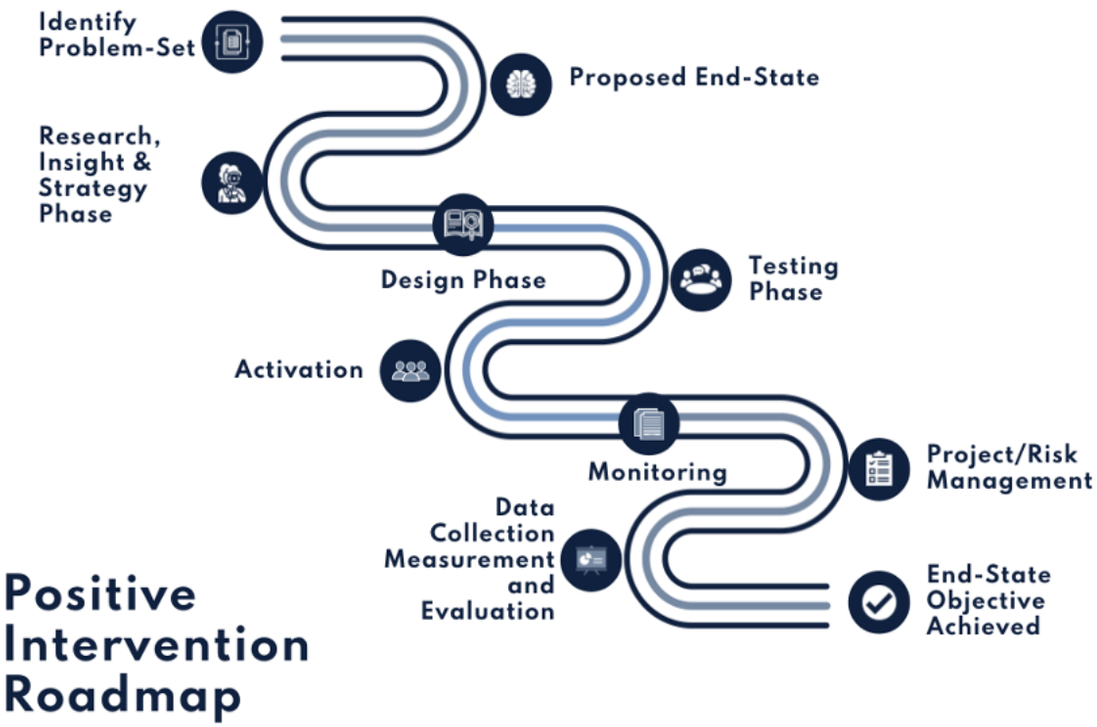
The presence of the term resilience building in the online intervention space becomes more of a long-term success criterion or factor when it is part of a drive to empower audiences to adopt peer support, mentoring, and self-reliance narratives alongside having prebunking, counter, or alternative examples to lean on. These processes not only have positive effects on engagement and trust but also build self-protection and community self-regulation and management into the mindset of audiences who may be very keen on doing more and being part of something bigger. This approach directly fosters both resilience and empowerment alongside offering tangible measurement opportunities for evaluation purposes.

As an obvious contrast, resilience in the offline context is something that requires building, constructing, and layering support, capacity building, resources, partnerships, and infrastructure over a sustained period of time. As such, it is good practice to view this approach as distinct to direct counter-narrative or counter-extremism efforts, where there is an identified need to address deep-rooted ideological issues manifested within

communities. This is where the role of aspiration as a form of incentivization within online resilience-building efforts can be useful to keep audiences engaged over time.

The tactics, considerations, and challenges of deploying these various methods highlight the need for positive interventions to be based on process and insight-driven stages, phases, and pathways. The more in-depth the insight and planning processes are, the greater the likelihood of the creative and content dissemination phases will yield positive results for the overall project's objectives/end-state needs. To illustrate this point, the roadmap below amalgamates the processes and phases that are key components of an effective positive intervention:

Figure 6: Positive Interventions Process



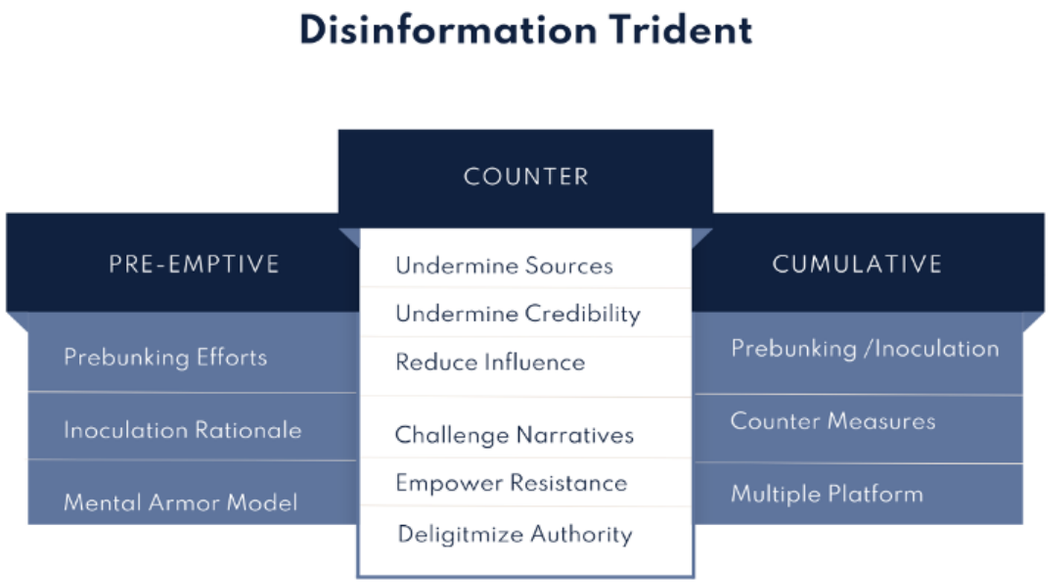
What Can We Learn from the Disinformation Space?

Extremist narratives and disinformation are surprisingly similar from the point of view of having a symbiotic relationship. It can be argued that extremism is a branch and disinformation its root, in that both need each other to thrive and evolve. In terms of the online space, different emergent and historical factors come into focus regarding the role of tech and civil society that require discussion and contextual clarity in terms of learning, best practices, and pitfalls. An obvious place to start is with the assertion that disinformation is not new, and like extremism has morphed to mean different things at different times in history. What is new is that the proliferation, scale, and size of disinformation are at unprecedented levels and permeate all aspects of social and digital life.

Online disinformation is not limited to one subject matter or context. Advances in technology and the use of tech tools have made huge inroads into tackling disinformation, misinformation, and propaganda through data

mining, machine learning, and advanced algorithmic capabilities. The issues remain those of scalability and volume of response against scale and dissemination of disinformation. Experts maintain that information is the new weapon of choice for harmful actors and VEOs. Terminology has appeared in recent times that suggests this issue is taking up more and more thinking space for government, tech sector, and civil society. Terms like "narrative warfare," "durable disorder" and "weaponized words" are just some examples of how ubiquitous the issue of disinformation has become.[56] What can the CVE sector learn from approaches to tackling disinformation and other forms of harmful information online? Essentially, this battle for information influence and the weaponizing of narratives has prompted tech-centric actors to initiate the use of tools to identify, target, take down, refute, and challenge disinformation through a more direct relationship with end-users. What this translates into looks like the following trident:

Figure 7: Disinformation Trident



Pre-emptive counter-disinformation involves the use of take downs, censorship, and removal from platforms (account suspension). Active counter-disinformation involves using algorithms and machine learning to warn and engage end-users when they are interacting with disinformation, fake news, or alike, creating a choice dilemma for the end-user and offering the chance to learn more about the threat posed by disinformation. Cumulative counter-disinformation is based on taking fact-checking, critical thinking, rational choice modeling techniques, and offering alternative sources of information by redirecting end-users to other places by treating them with an empathetic manner to build trust and maintain it.

A key lesson from the disinformation space is that it is always morphing and oscillating on different wavelengths in order to attract and then retain attention. Disinformation plays a "long game," basing itself on a strategic

56 Caroline Jack, "Lexicon of Lies: Terms for Problematic Information," Data Society, August 9, 2017, https://datasociety.net/library/lexicon-of-lies/.

plateau that seeks to build up momentum over time in order to undermine rational thinking, objectivity, and trust in traditional modes of authority with one overall end-state in mind – the person becoming the disinformation. These different wavelengths refer to the use of different means, mediums, and technologies to confer similar meanings but with subtle variations. Examples include memes, humor, satire, deepfakes, animations, trolling, and friction. These may all have different visual facets but carry the same underlying narrative on whatever the subject matter is. The aim here is to attract users with at least one of these "hooks" in order to then develop a longer-term relationship to a point where this user is now actively creating, sharing, and adhering to the desired narrative.[57] The use of "alternative facts," canceling, and trolling techniques using bots start to create a chasm in the thinking space of the end-user, which can often lead to cognitive dissonance. Another learning from this context is that disinformation is very much a mind-game pursuit in its initial stages, as it seeks to replace rational thinking and emotional intelligence with conspiratorial ideas and groupthink.[58] Extremism in some guises also adopts a similar tactic through narratives of guilt, emotional cues, sentiment generation, and redemption arcs, eventually creating a "them and us" mindset. Some relevant case-study examples that address the nexus between disinformation and extremism can be found in the work by Moonshot CVE (see below).

Two recent projects aim to counter disinformation, political polarization, and violent extremism by encouraging internet users to think more critically about the information they consume. These campaigns prompted internet users to consider how their behavior affected their online peers and offered detailed resources or one-to-one support on how to hold constructive conversations that bridge ideological and political divides. This section provides an overview of these projects and how they engaged online audiences beyond delivering a counter message or alternative content.

Using Gamification to Advance Media Literacy in Indonesia

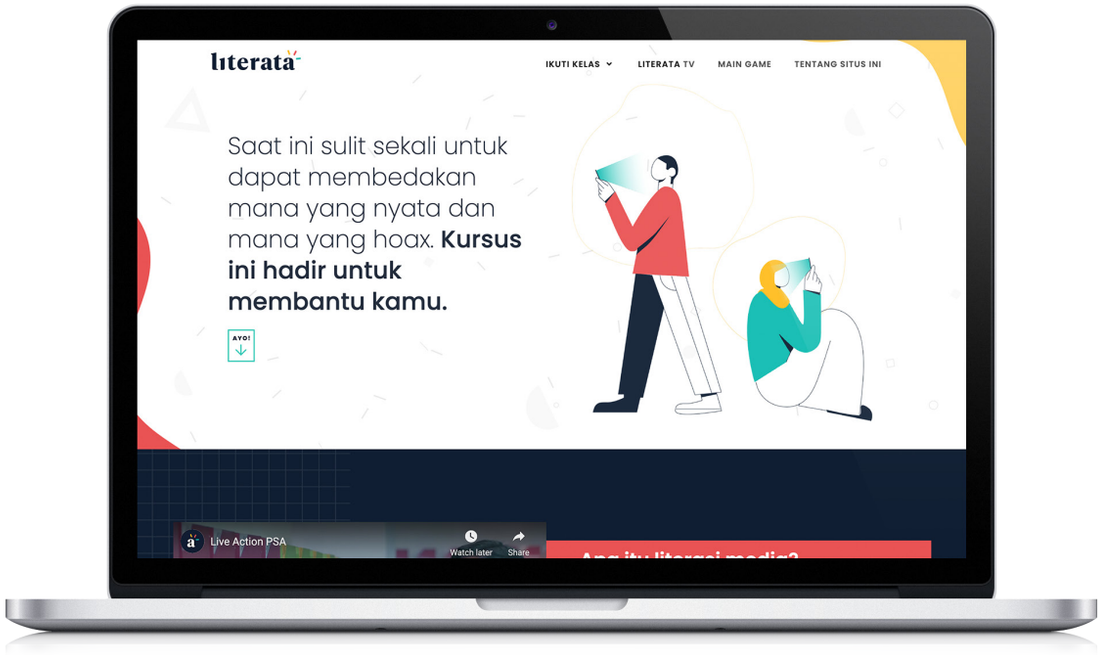
Moonshot has monitored the disinformation environment in Indonesia since 2019 and developed a database of key disinformation narratives and the audiences consuming them. In partnership with the University of Notre Dame, the International Research & Exchanges Board (IREX), and GeoPoll, they used this research to deploy digital campaigns and reach vulnerable audiences with a media literacy website and gamified content.

The website, Literata.id, is designed to guide new digital arrivals who encounter disinformation on the internet and improve their media literacy skills. Literata.id contains eight lessons from the IREX curriculum, including how to spot echo chambers, how the language used in news articles can mislead readers, and how to identify different types of disinformation. Each lesson includes a video and a short quiz to test users' engagement and understanding of the content.

57 "Prebunking Anti-Vaccine Narratives: An Effective Alternative to Debunking Individual False Claims," Jigsaw, March 2, 2022, https://medium.com/jigsaw/prebunking-anti-vaccine-narratives-an-effective-alternative-to-debunking-individual-false-claims-78f0047a8b47.

58 "Hate Clusters Spread Disinformation Across Social Media. Mapping Their Networks Could Disrupt Their Reach," Jigsaw, July 28, 2018, https://medium.com/jigsaw/hate-clusters-spread-disinformation-across-social-media-995196515ca5.

Figure 8: The Literata.id website



Between September 2020 and April 2021, Moonshot developed and tested a media literacy game called Gali Fakta designed for the same audience. The script transposed lessons from the IREX curriculum and was built around real examples of disinformation from Indonesia.

The decision to use real disinformation in the game was informed by the inoculation method developed by researchers at the University of Cambridge. The method is based on the theory that psychological resistance to disinformation can be developed by exposing individuals to weakened versions of fake or manipulated stories that they will come across in the real world. Moonshot exposed users who had previously engaged with disinformation with select examples in prepared and carefully managed settings. All examples included a disinformation warning or a rapid feedback loop so that players were clear on what was true and what was false. Moonshot also created a second version of the game with low-risk examples of disinformation for the purposes of general consumption.

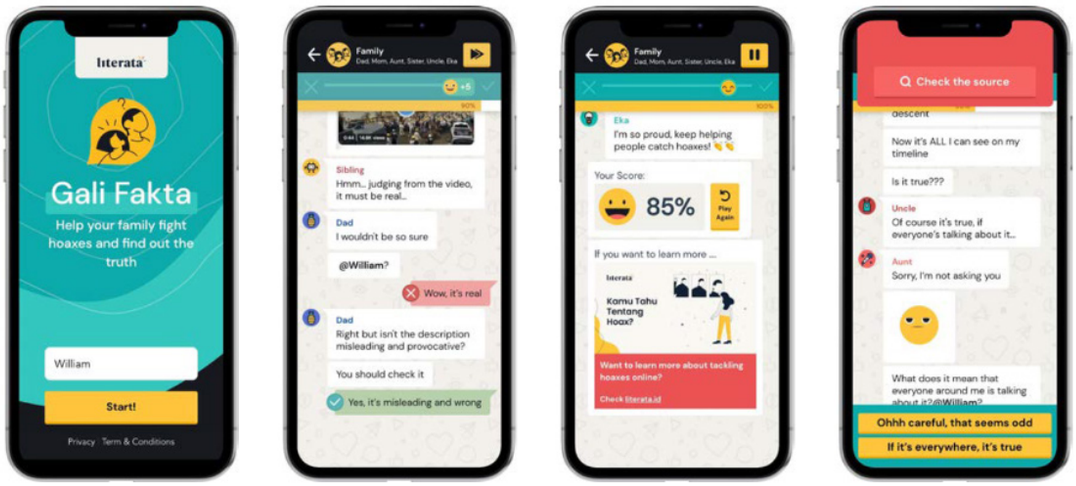
How the game worked for users

Gali Fakta is designed in the style of a family WhatsApp chat. This familiar and entertaining scenario reflects the real-world context of how disinformation spreads in Indonesia, where WhatsApp is the most popular messaging application. Family chat groups make up over 70% of Indonesian user activity on the platform. A similar approach can be seen in MediaSmarts work focused on youth-focused programming.[59]

Figure 9: Gali Fakta Inoculation Game

59 "Reality Check: The Game," Media Smarts: Canada's Centre for Digital and Media Literacy, n.d., https://mediasmarts.ca/digital-media-literacy/educational-games/reality-check-game.

Gali Fakta: The disinformation inoculation game



The game teaches IREX's media literacy lessons through leading questions and social proof. When a user is prompted to spot disinformation, correct answers are rewarded by points or the family reaching a consensus. Incorrect answers are docked points and met with general confusion by family members. The player's "cousin," Eka, is a media literacy expert and functions as a corrective voice should the player get any answers wrong. In addition to Eka stepping in to correct them, players who answer incorrectly are also immediately given the chance to correct themselves. Should they stick to their original incorrect answer, they lose more points and receive an explanation from Eka.

Between April and October 2021, Moonshot ran campaigns on Twitter, Google Search, and Google Display to reach audiences in Indonesia. They simultaneously advertised the website and game, and when these campaigns were complete, Moonshot used post-surveys to measure whether engagement with the media literacy content had impacted participants' self-reported intentions to respond responsibly and proactively to online disinformation.

How users interacted with the media literacy content

Overall, 24,581 individuals visited the website, of whom 72 took a quiz and 289 watched a video. 8,128 individuals viewed the game page, of whom 781 started playing it and 98 completed it. The website had more visitors and a lower bounce rate than the game. On the website, users could access content right away, whereas the game requires the user to type in a username and commit "10 minutes of their time."

Despite the website's lower bounce rate, the game was more effective at engaging users and maintaining their attention. Game players spent approximately 12 times (1,185%) as long engaging with the media literacy content than those who visited the website without playing the game. Results were statistically significant at the 95% confidence level ($p < .05$).

However, the impact of media literacy content on behavior change was inconclusive. Differences between control and treatment groups taking the survey made it impossible to measure whether or not there was a statistically significant behavior change, as the control group members self-reported as being significantly younger and more highly educated. When these demographics were controlled for, the sample size was too small to conclude if the game increased players' media literacy.

How this approach can be improved

This outcome raises a number of considerations for future programming. A key lesson is the need to more robustly connect user behavior to survey responses. Moonshot's design did not connect data on user behavior from Google Analytics with the users who responded to surveys. Attributing multiple behaviors to a persistent user in this way is challenging, but the stronger the link between user behavior and the results of their post-treatment survey, the more likely a program is to confidently measure any changes in behavior.

The program has been extended for four years to enable further adaptation and testing. Moonshot is exploring a number of possible solutions with its partners, including new survey software which would enable pseudonymous connections between user behavior and pre and post-survey results, and collaborating with GeoPoll to SMS or call users after their treatment.

Preventing U.S. Election Violence Through Strategic Messaging

In November 2020, Moonshot launched a new model for election violence prevention and peacebuilding. Over four months, the campaign sought to counter and reduce the threat of violence in the aftermath of the 2020 U.S. presidential election by safeguarding internet users seeking or consuming content that could incite violence and offering alternative pathways that supported shared values, inclusive citizenship, and mental health.

How the campaign was designed

Key audiences included individuals engaging with QAnon, white nationalism, violent armed groups such as the Proud Boys and Oath Keepers, and disinformation related to the results of the 2020 presidential election. The campaigns drew more than 1.7 million engagements from at-risk internet users, including retweets, clicks, video views, and downloads.

With input from local community partners, Moonshot designed and tested 21 pieces of original content containing de-escalation messages and promoting shared values, and redirected thousands of individuals to crisis counseling services. The campaigns' messaging, content, and targeting were refreshed continuously, based on an open-source analysis of post-election threats and partner consultations. Moonshot also surged de-escalation advertising at critical moments, such as the January 6 attack on the Capitol.

Moonshot's campaigns ran on four advertising networks: Google's Search, Display, and YouTube platforms, as well as Twitter. Beyond measuring standard indicators of audience engagement, such as ad impressions, views, and clicks, Moonshot evaluated individuals' engagement with partner websites, providing mental health support, guides on having difficult conversations with peers and loved ones, and promoting active listening skills.

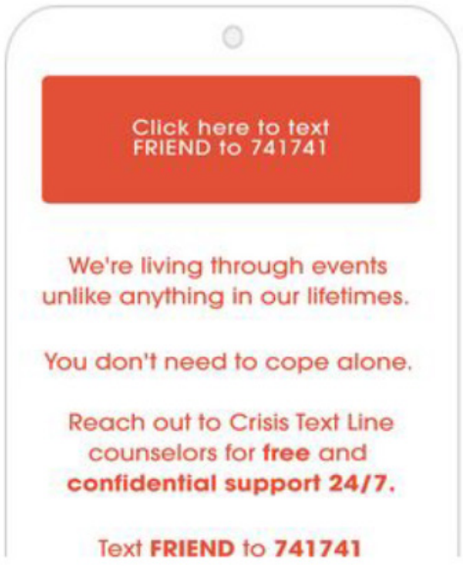
Impact and reach

Over 22,000 users engaged with an ad offering psychosocial support and visited websites featuring mental health resources or the Crisis Text Line (CTL) service. Thirty-three individuals texted the CTL helpline using a unique referral keyword, resulting in a total of 39 conversations, and it is likely that a greater number reached out using the generic keyword on the service's primary website (see Figure 10). This outcome provided evidence that connecting at-risk individuals with mental health support services can result in sustained engagement. To date, 163 conversations have occurred through Moonshot's partnership with CTL.

5,500 users visited a website focused on election integrity. The website was advertised via messaging about the importance of having healthy civic conversations with Americans holding opposing points of view. Over 100 visitors downloaded and shared bespoke discussion guides with titles like "How to Talk to Someone You Disagree With" and "How to Talk about the Election and Move Forward." Visitors downloaded 130 PDF guides in total. This provided evidence that Moonshot's audiences actively viewed and saved resources that fostered positive civic engagement and challenged polarization and hate. Audiences interested in election disinformation were most likely to visit the website and downloaded the most content.

1,250 users visited a website featuring actionable steps on how to process difficult emotions and build a supportive local community. This website was advertised via messages about the importance of neighborliness and inclusive American identity. 18 individuals interacted with sections of a "Get Involved" page, which provided information on how to process and express emotions in a healthy way, how to care for themselves and other individuals, and how to solve challenges collectively. Visitors also reviewed a "conversation prompts" page, providing information on how to have productive conversations with other Americans.

Figure 10: Crisis Line Microsite. Visitors who clicked the red button were instantly redirected to their preferred messaging application to start a conversation with a counselor.



Overall, Moonshot found that offering psychosocial support is a uniquely effective way to engage audiences at risk of participating in political violence. While messages were at times met with skepticism and occasional hostility, more often content was engaged with positively, and shared and endorsed enthusiastically. Through the partnership with CTL, Moonshot and its partners were able to facilitate conversations with crisis counselors for 33 Americans, suggesting that members of these audiences are indeed open to repeated engagement and may continue to use psychosocial resources after initial exposure. This also indicated the value of involving mental health organizations in future projects.

The difficulty in directly applying learning from the disinformation space to the CVE space is that where race, religion, and creed/doctrine-based narratives are present, the skill in using tech tools, creating warnings, and building cumulative resistance to harmful ideas becomes fraught with confusion over representation, authority, and credibility. However, as findings from pilot testing in prebunking approaches begin to emerge and help to shape future plans, machine learning capability continues to gather at pace, and a greater effort is made to align the private-tech sector with the CSO sector, is it reasonable to expect a much larger footprint of pre-emptive, active and cumulative actions to become commonplace within online CVE efforts. The age of information warfare brings with it the possibility that merely winning the narrative battle to be first or "own the narrative" will not be enough, because without discrediting and fully undermining the opposing narrative, ideas endure simply because they can now be repurposed to become relevant again and more durable. The emerging set of considerations seems to be how to work at a mass scale in the pre-emptive space while simultaneously ensuring the micro-targeting tools are fit for purpose from a moral, ethical, and impact standpoint.

Conclusion

The convergence and continued growth of narratives that espouse binary and otherization facets within their central ideas mean 'polarization' and 'unity' alike are now tools more easily weaponized through communication engagement and lifestyle choices. Whereas CVE has morphed into its own industry with accepted norms and practices, VEOs, terrorist networks, and their supporters have often been able to harness technology and engagement to their advantage, while the policy, tech, and civil society sectors often struggle with how best to use these same tools within the confines of legal, social, and ethical considerations.

This report sets out key learning on central issues linked to effective strategic communication in the online space, and in doing offers both opportunities and challenges to the sector. The presumed rise of disinformation (which has always been present and will always be) takes the focus away from how this tactic is used by extremist groups and instead concentrates too heavily on its ontological roots. Visibility of disinformation is a distinct issue to the use of disinformation, as is the presence of extremism in relation to the proliferation of extremism. The prebunking approach can make huge strides in not only tackling conspiracy narratives but in building more effective prevention inputs for the CVE space. The use of influencers within CVE has become meshed alongside the ongoing debates around using "credible" voices, and this context needs further research and some form of conceptual framework development in order to be better understood.

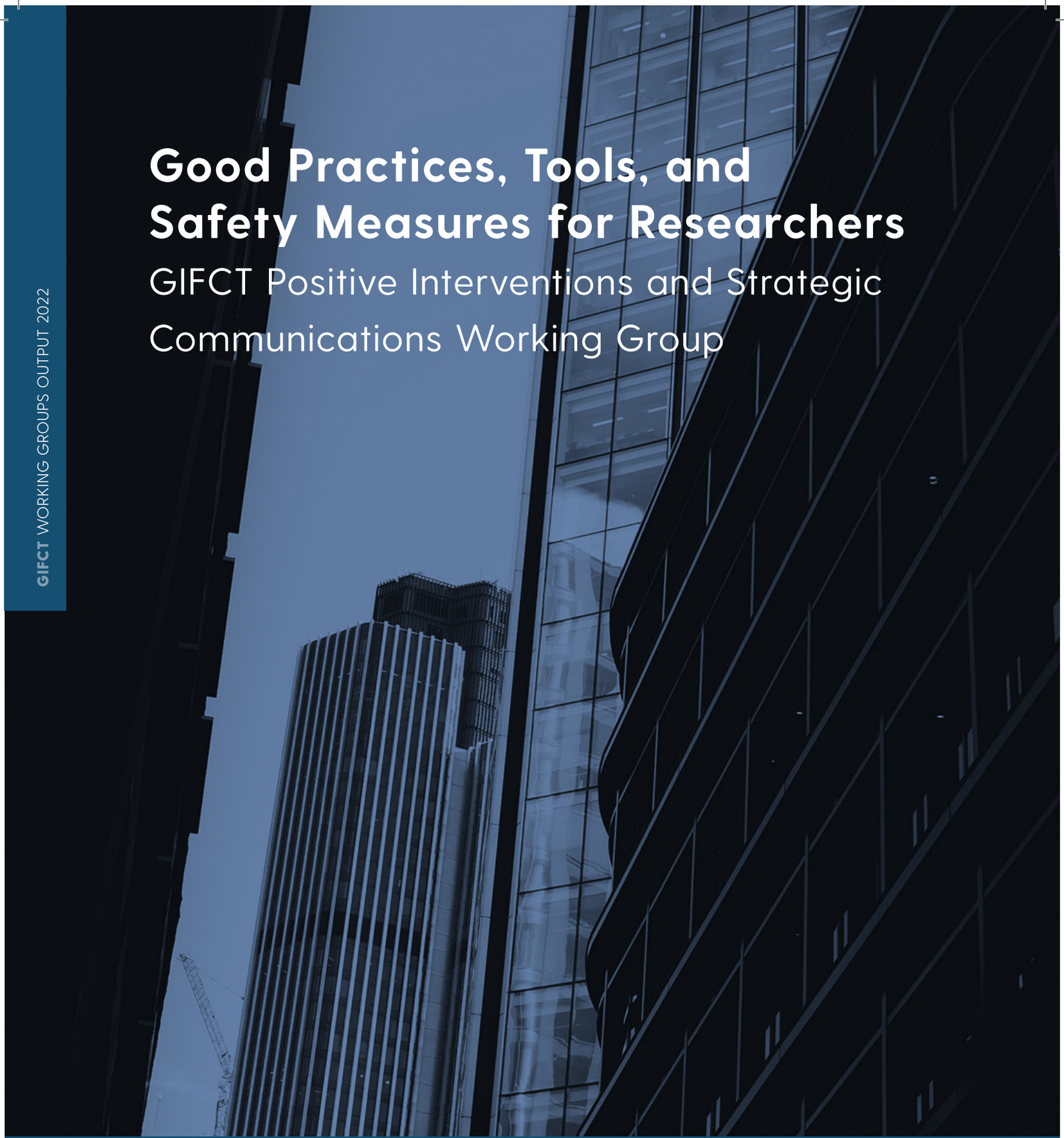
The strategic and tactical considerations highlighted in this report outline the need for more emphasis to be placed on understanding audience behavior(s) better and at a more granular level. The relationship between problem sets, strategy, and end-state objectives must be more closely aligned to the use of innovation, offline

engagement learning, and diversity in tactical approaches (e.g., humor, incentives, and friction) that have the potential to enhance content as well as engagement. These issues relate at the macro-level to the continued need for measurement and impact assessments to be based on a more nuanced understanding of the role sentiment, semiotics, individual eco-systems, and legacy impact on both campaigns and their recipients.
The presence of diverse and often hidden harmful actors in the online space means the tools offered by the tech sector are the key to undermining hate, extremism, and violent extremism, but also point to the need for increased private-public partnerships to sit at the heart of CVE efforts. This report has highlighted that a key tool for this sector to effectively counter violent extremism is the relationship basis between those skilled in the practice of CVE and those equipped with the resources to help deliver results.

GIFCT WORKING GROUPS OUTPUT 2022

Good Practices, Tools, and Safety Measures for Researchers

GIFCT Positive Interventions and Strategic Communications Working Group



GIFCT
Global Internet Forum to Counter Terrorism

Kesa White
Polarization and Extremism Research and Innovation Lab (PERIL)

Introduction

When collecting data on actors such as terrorists and extremists, researchers (along with content moderators and investigators) are required to put themselves in harm's way for the greater good of society. A primary researcher investigating extremist activity online has one of the most vital and dangerous positions at a research organization, academic institution, or tech platform. The researchers in this area serve a vital need because they are engaging with actors online, which involves putting themselves at risk.[1] Similarly, content moderators (along with trust and safety specialists) tasked with reviewing dangerous content face similar (and in some cases even more severe) exposure to harmful data in the course of their work.

While the safeguarding of researchers should be a consideration of any research project, the sensitivity of extremism and terrorism research, along with the potential harms that researchers face, place a special onus on organizations to ensure researcher wellbeing. Conducting research for academic and organizational purposes has its own set of unique risks that lack a comprehensive response that would alleviate potential harms, including exposure to upsetting or potentially traumatic content, impacts on mental health, or in extreme cases risks to personal security.[2] Due consideration should be made of potential impacts on researcher welfare and sufficient mitigating actions should be enacted. Although a researcher may not be in a physical conflict zone, advanced security measures cannot always defend against bad actors who are constantly discovering new and innovative methods to harm individuals. This paper aims to help mitigate risks researchers may encounter by offering solutions and resources to assist them.

1 Maura Conway, "Online Extremism and Terrorism Researcher Security and Privacy: Some Practical Advice," Global Network on Extremism and Technology (blog), February 19, 2021, https://gnet-research.org/2021/02/19/online-extremism-and-terrorism-researcher-security-and-privacy-some-practical-advice/.

2 Miriah Steiger et al., "The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support," in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21) (New York, NY: Association for Computing Machinery), Article 341: 1–14, https://doi.org/10.1145/3411764.3445092.

Types of Threats

Threats of any kind should be taken seriously because it is impossible to recognize whether a threat may result in actual violence or other harm. The most common threat categories that researchers and at-risk audiences (individuals such as journalists, activists, and politicians who are frequently targeted with harassment) are exposed to include direct, indirect, veiled, and secondary:[3]

- **Direct:** A specific target and victim are explicitly identified by name. All threats of this nature should be taken particularly seriously as the named victim may be at risk of potential violence or attack.
- **Indirect:** Denoted by vagueness, lack of clarity, and non-specifics of a violent attack which do not provide enough information to understand the author's intentions.
- **Veiled:** The threat does not specifically call for violence because it is vaguely implied. The rhetoric used in this category of threats is often observed in online behaviors such as shitposting and trolling.[4]
- **Secondary:** These threats are often not from a specific actor and may include exposure to harmful content, the experience of vicarious or secondary trauma due to the nature of the content, interviews, or narratives in the research, and other potential threats posed by (for example) failing to protect from exposure the Personally Identifiable Information (PII) of the researcher or organization.

These threat categories represent only a few examples of potential techniques actors can use against researchers and at-risk groups.

Risk Management Frameworks for Researchers

The principles below provide a general risk management framework designed to be flexible for individual researchers, institutions, and specific populations and identities who may be at additional risk.

Do No Harm

A core principle of applied social research, especially on sensitive topics such as extremism, terrorism, and conflict is that of "do no harm."[5] This fundamental tenet is derived from the humanitarian and medical sectors and requires that researchers prevent and mitigate against harming study participants and their wider

3 University of Arkansas Little Rock, "Types of Threats," University Police, n.d., https://ualr.edu/safety/home/emergency-management-plan/threat-assessment-team/types-of-threats/.

4 Note: Shitposting refers to off-topic, aggressive, and frequently ironic online posts often made in a public or semi-public group for the purpose of derailing or disrupting conversation or to upset others. Similarly, trolling is an online practice through which the instigator (the 'troll') engages in arguments of an illogical, spurious, or directly debasing nature. Trolls and shitposters typically use anonymous or pseudonymous handles to mask their identity. The actual intention of the troll or shitposter may be a genuine if misguided attempt at humor or credibility-building (doing something for 'the lolz').

5 The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, "The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects for Research," April 18, 1979, https://www.hhs.gov/ohrp/sites/default/files/the-belmont-report-508c_FINAL.pdf.

community (whether intentionally or out of negligence).[6] Additionally, researchers and the institutions backing them should ensure that staff is not harmed through their work and that risks are mitigated to the highest extent possible. This principle should inform all research efforts in the terrorism and extremism research sphere.

Two corollary principles, also from humanitarian work, are that staff both have the right to cease their work at any time if they are in danger or are at a personal limit and that they have no "right to remain" after the organization ceases a project to protect its staff. In research, this implies that researchers ought never to be coerced to continue research when facing personal threats or having reached their limit[7] and that individual researchers inside an organization must not continue working on a project if (after a risk management assessment by the institution) the risks are deemed too high to pursue completion.

Research Institutions: Duty of Care and Risk Management

Research institutions have a primary responsibility to protect themselves and their researchers from a variety of harms, including traumatization, harassment, doxing threats, and intimidation. Due to researchers investigating and studying a variety of topics that have real-world dangers, research institutions are responsible for protecting the well-being of their researchers from potential harm.[8] The level of funding, institution staff size, and other resources can limit the amount of protection they are able to provide. However, this should not be an excuse for not providing a minimal level of risk management that can be facilitated regardless of capacity constraints. If an institution is unable or unwilling to commit to a duty of care to its staff on a project, a serious ethical and pragmatic evaluation of undertaking that project should be held.

A key element of the duty of care approach is ensuring that prior to the commencement and initiation of any fieldwork (online or offline), a thorough and holistic risk assessment and management plan are created.[9] The risk assessment should include the likelihood of certain outcomes occurring that may cause harm, damage, or threat to the researcher, institutions, and (where relevant) the research subjects. Once this assessment is completed, a standardized numeric or threat level indicator system (for example, low-medium-high) should be instituted. This risk indicator system (the complexity of which again depends on project size, scale, and scope) should then be further developed by clear and process-driven mitigation and management-based actions that clearly outline the identification, actioning of decisions, and ongoing monitoring of risks (should they materialize).[10] This relates to the chain of command, practical steps, and naming those responsible for administering each step and process (accepting that external agencies may also be part of this stage of risk management). A risk management plan works best when it is part of an overall project management system that includes distinct

6 Won Oak Kim, "Institutional Review Board (IRB) and Ethical Issues in Clinical Research," *Korean Journal of Anesthesiology* 62, no. 1 (January 25, 2012): 3 – 12, https://doi.org/10.4097/kjae.2012.62.1.3.

7 The Sphere Project, "Humanitarian Charter and Minimum Standards in Disaster Response," United Nations High Commissioner for Refugees, August 29, 2001, https://www.unhcr.org/en-us/partners/guides/3b9cc1144/humanitarian-charter-minimum-standards-disaster-response-courtesy-sphere.html.

8 Alice Marwick, Lindsay Blackwell, & Katherine Lo, "Best Practices for Conducting Risky Research and Protecting Yourself from Online Harassment (Data and Society Guide)," (New York: Data & Society Research Institute), 2016, https://datasociety.net/wp-content/uploads/2016/10/Best_Practices_for_Conducting_Risky_Research-Oct-2016.pdf.

9 Paul Hopkin, Fundamentals of Risk Management: Understanding, Evaluating and Implementing Effective Risk Management (London: Kogan Page Limited).

10 The United Nations Office for Disaster Risk Reduction, "Towards the Post-2015 Framework for Disaster Risk Reduction Indicators of Success: A New System of Indicators to Measure Progress in Disaster Risk Management," Prevention Web, November 21, 2013, https://www.preventionweb.net/files/35716_newsystemofprogressindicatorsfordrr.pdf.

stages, controls, and processes for every aspect of research, communication, and day-to-day decisions. The creation of a risk log allows for a named person to manage and monitor duty of care issues.

It is vital for institutions to be aware of the dangers researchers face to develop precautionary measures. When harm against researchers and institutions shows signs of potential violence and hampers individual wellbeing, legal action should be taken and law enforcement should be informed of the threat. All threats should be archived to monitor the actions of the perpetrator because online risks can grow into real-life actions. However, falling short of these most extreme examples, a methodological and institutional policy should be implemented to safeguard researchers.

At a minimum, duty of care should include understanding the risks and developing prevention mechanisms for:

- Exposure to harmful and toxic content, including the potential for vicarious or secondary trauma responses among researchers;
- Virtual harassment or threats from target research populations, including violent extremist organizations and individuals;[11]
- Physical threats or attacks against researchers or institutions; and
- Legal obligations and threats in the specific context of the researcher, as data collection may expose researchers to downloading or receiving terrorist or extremist content for which legal penalties may apply depending on the jurisdiction.

While no institutional approach is perfect in safeguarding staff all the time, key principles used by research organizations in this space include:

- Providing non-coercive, "opt-in" policies for research staff to ensure they are not forced to work on efforts they may find particularly upsetting or traumatizing. This could include briefing analysts on the potential risks to their mental wellbeing at the outset of a research project and the signing of consent forms for analysts;
- Providing accessible and effective mental health and wellbeing support (ideally in tandem with certified medical professionals) through providing resources and access to compensated time-off for staff;[12]
- Directly de-briefing with researchers on harmful content and exposure to challenging material;
- The training of project managers on best practices for discussing mental health issues with staff;[13]
- Ensuring efforts are made to ensure that individuals affiliated with a research institution but not directly working on harmful content are not accidentally exposed to traumatic material;
- Ensuring there is a basic level of data and information security (infosec) at the individual and organizational level to protect against data breaches, network infiltration, and inadvertent exposure at the device to network levels;
- The use of encrypted communications whenever possible, including through free tools like Signal, ProtonMail, VeraCrypt, and vetted Virtual Private Network (VPN) services;

11 "So You've Been Doxed," Crash Course Override Network, n.d., http://www.crashoverridenetwork.com/soyouvebeendoxed.html.

12 "The Importance of Providing Mental Health Days: They're a Valid Reason to Stay Home," Wellmark, n.d., https://www.wellmark.com/blue-at-work/healthy-employees/provide-mental-health-days.

13 "Why You Need to Talk About Mental Health in The Workplace: Five Tips for Addressing It," Wellmark, May 2020, https://www.wellmark.com/blue-at-work/healthy-employees/mental-health-in-the-workplace.

- Reviewing legal guidelines in the jurisdiction of the researchers and ensuring compliance or safeguarding against potential infringement; and
- Developing and deploying sufficient physical and virtual staff safety protocols, such as reporting lines for threats, pre-established communication links with law enforcement, and mitigation tactics such as identity monitoring, PII removal services, and physical security hardening through surveillance systems, locked and secure offices, and (if needed) security for researchers.

Malicious actors can target an institution and researcher at the same time by leaving messages on an institution's answering machine regarding the targeted researcher. These types of messages put the researcher and the institution at risk because individuals can increase their level of action by showing up to the institution in person creating a threatening environment.

Women and Minority Researchers

If they encounter extremists, researchers who identify as women will likely experience power dynamics, sexual harassment, and dehumanization aimed at belittling them. Extremists expressing misogynistic views may attempt to commit physical or virtual violence against female researchers. Similarly, individuals identifying with LGBTQ+ communities, especially those who express their sexual orientation and/or gender expression in particularly identifiable ways, may find themselves facing additional risks than other researchers. Mitigation measures, especially in research carried out virtually, can be effective for protecting both women and LGBTQ+ identifying people, whose diverse views and analytical intelligence are necessary to include from a research perspective.

The threats against researchers and at-risk audiences are often the result of a publication, media appearances, or random acts of trolling. Many extremist individuals, organizations, and groups have a history of using violence and symbolism to reach their goals, which demonstrates the threat they can pose offline in addition to their virtual threats. As researchers become more accessible and build a public profile with their work – such as by publishing or appearing in media interviews – the opportunity for malicious actors to target them increases. The risks researchers face may hinder their ability to safely interview certain populations and individuals, implying a need to mitigate interviewing risks based on the danger posed by the target interviewee based and the identity expression(s) of the researcher.[14]

14 Irène Bahati, "The Challenges Facing Female Researchers in Conflict Settings," Governance in Conflict Network, June 10, 2019, https://www.gicnetwork.be/bukavu-series-the-challenges-facing-female-researchers-in-conflict-settings/.

to a narrative of "white genocide," the white race in America is at risk of being wiped out as minorities are entering the country at an alarming rate. To white supremacists, being a minority and an outsider to their community automatically makes that researcher a threat, as they are supposedly contributing to wiping out the white race.[15]

The dangers of conducting research on extremism, terrorism, and violence can be reduced by taking measures to protect oneself online and offline. Personal protective measures can include using a secure virtual private network (VPN), paid or self-led services to remove personally identifiable information online, maintaining additional vigilance when publishing or publicly discussing research, and robust online hygiene and infosec practices (links to useful toolkits in this regard are provided in the annex to this report).

In-person and virtual interviews can also benefit from the added security and changed power dynamics provided by conducting interviews with a colleague, especially one who may be able to mitigate social biases from the interviewee. One helpful practice can be paired interviews with male and female colleagues, younger and older, or LGBTQ+ groupings (evaluated on a case-by-case basis). Utilizing safe risk mitigation practices decreases female and minority vulnerability to malicious actors.

Directional Harms

The Obligation to Report

In this space, researchers will likely come across and may directly be analyzing, harmful content which violates platform-specific 'terms of service' (TOS) policies, along with content that may be outright illegal. Furthermore, researchers may also unearth specific credible threats to others, which carry an additional set of risks.

In general, researchers must use their best judgment when deciding whether to report problematic content encountered in the course of their work.[16] Researchers often assume the content being reported will be removed from the platform, which could hinder their ability to access valuable information. However, content that may indicate a user or actor likely poses a direct threat to themselves or others should be treated with extreme caution and in many cases ought to be reported to the platform hosting the content, law enforcement, and other relevant authorities. Researchers and their institutions should be familiar with local legal requirements as these vary: in some countries, researchers are protected when downloading terrorist or potentially illegal content in the course of their work. However, in other contexts, even accessing such content for the purposes of empirical research may be illegal. Similarly, the requirements to report content to government actors may vary

15 "White Genocide," Anti-Defamation League, April 5, 2017, https://www.adl.org/resources/glossary-terms/white-genocide; Milan Obaidi et al., "The 'Great Replacement' Conspiracy: How the Perceived Ousting of Whites Can Evoke Violent Extremism and Islamophobia," *Group Processes & Intergroup Relations* (August 2021), https://doi.org/10.1177/13684302211028293; Manuel Castro e Almeida & Alistair Harris, "The Conflict Sensitivity Principle: Can Best Practice in Conflict Research Fill the Ethics Gap in Terrorism and Counterterrorism Research Practice?," *Terrorism and Political Violence* 33, no. 2 (March 24, 2021): 381–396, https://doi-org.proxyau.wrlc.org/10.1080/09546553.2021.1880159.

16 J.M. Berger, "Researching Violent Extremism: The State of Play", Resolve Network, June 2019, https://resolvenet.org/system/files/2019-09/RSVE_RVESeries_ResearchingViolentExtremism-TheStateofPlay_JMBerger_June2019.pdf.

depending on location.

TOS and content policies vary substantially from platform to platform. For example, content that goes against Twitter's policies may be acceptable behavior on Gab, so a researcher must use their best and informed judgment when researching across platforms. Most platforms, and Global Internet Forum to Counter Terrorism (GIFCT) members especially, have some extent of public-facing community or content standards for what is and is not allowed on that platform. Additionally, most platforms have a way to report problematic or harmful content. Ultimately, most platforms employ their own trust and safety teams responsible for investigating and reporting online behaviors, meaning that researchers do not have the sole responsibility to report content, nor will every piece of reported content ultimately be removed or acted on based on platform policies.

In general, ethical practice in social science and social work dictates that the anonymity of research participants may be lifted in cases where the participant poses a credible threat to themselves or others. A similar approach can be taken in extremism and terrorism research. Violent extremists have established an active presence on social media prior to their attacks to post content and interact with material related to their ideology. If a researcher encounters a threat of imminent, credible action that would cause extreme harm (e.g., a planned terrorist attack or school shooting or an identified user credibly threatening suicide), reporting to authorities should be strongly considered.

While a researcher is examining extremist content online, they noticed that some of the content posted contained racist rhetoric. However, it did not necessarily need to be reported to the social media platform per content standards at the time. While racist material is hurtful, reporting and focusing on every instance of racism can cause moderators or researchers to overlook more serious calls for offline violence.

Protection from Platforms

Social media platforms have a limited duty to protect researchers and other users from online harm that occurs on their websites. The policies and regulations enforced by companies use a variety of tactics to address violations on their platforms. For example, Twitter offers "account-level enforcement" and "direct message-level enforcement" depending on a user's violation.[17] The enforcement can include stopping a user's messages or placing an account in "read-only" mode which prevents a user from tweeting. Some platforms may also (if asked) ensure researcher and user information is protected and cannot be accessed without the user's permission. In general, firms have the same responsibility to protect researchers as they do with protecting their

17 "Our Range of Enforcement Options," Twitter Help Center, n.d., https://help.twitter.com/en/rules-and-policies/enforcement-options.

other users.

However, as research findings in this space can be highly sensitive, researchers may also experience blowback from platforms or publishers after releasing their work. Without any intention of incurring bias in research, as a mitigating step the investigating team or individual may wish to engage the platform in conversation about the research aims. Most platforms are interested in reducing harmful content hosted on their servers: asking for input on research questions may be mutually beneficial during the research design phase. Similarly, providing an opportunity for platforms to comment on findings may provide additional context prior to public release. Still, the profile of the platform should be considered when determining if such an engagement would be beneficial or fraught, especially when considering approaching platforms that are specifically created for use by extremists, have ties to authoritarian regimes, or have previously proved to be unwilling to engage on harmful content.

Law Enforcement

When conducting research into extremism and terrorism online, there is the very real and likely possibility that analysts encounter illegal activity. This could include networks of individuals sharing content produced by proscribed terrorist groups, evidence of individuals recruiting for terrorist movements, or even credible threats of a terrorist attack or hate crime. Additionally, researchers themselves might find themselves targeted by individuals as the result of publishing insights on extremist and terrorist content online through (for example) doxing and online harassment. Taking these possibilities into account, researchers should consider scoping out appropriate mechanisms for referring to law enforcement (such as online portals) at the outset of a research project.[18] However, the capacity, human rights standing, and types of response by authorities may, depending on the context, also be considered in the ethical call of whether and to whom a report should be filed. To facilitate this, in advance of commencing active fieldwork, research organizations should develop clear "bright lines" for reporting harms.

External/Audience Harms

Informed Consent

Informed consent is the groundwork of ethical human-subject-centered research. In short, if data collection is not carried out via 100 percent open, publicly accessible information and views, and researchers intend to speak with or observe participants in non-public spaces, informed consent must be obtained. For consent to be obtained, it must fulfill three criteria by being free (voluntary and able to be withdrawn at any time), specific (directly related to the research aims), and informed (whereby the participant is fully aware of potential consequences of their participation). Study participants should be given full disclosure of the potential risks of their involvement and care should be taken that they understand those risks. Their anonymity must be ensured or the circumstances under which they may be named fully elucidated. While a full discussion of informed consent is beyond the scope of this document (free and comprehensive guidance is readily available), researchers of

18 "Tell Us About Possible Terrorist Activity," UK Metropolitan Police, n.d., https://www.met.police.uk/tua/tell-us-about/ath/possible-terrorist-activity/.

online extremist and terrorist content must still adhere to these guiding principles in their work.[19]

However, gaining informed consent in online research is not clear-cut and perceptions of privacy online are complex. In some cases, online spaces are clearly public – such as a public Twitter feed – or clearly private, such as direct messages. However, in some cases, the privacy of some spaces is more ambiguous, such as in large open groups on Facebook, or in community forums where membership is required. Additionally, perceptions of what social media spaces are public and private may vary from the legal reality or the terms of service of a platform. These discrepancies between reality and perceptions of privacy are important, and accordingly, where information is gathered through a platform's Application Programming Interface (API) using (for example) a social listening tool, it is important to consider whether a research subject could reasonably perceive this information to be private.

Archiving Content

The online ecosystem is one of the most important sources of information for researchers to study communication, tactics, narratives, and other data on extremism. Websites and social media provide some of the richest data for researchers to analyze, which is why it is important for it to be properly archived for future research. At the same time, safely archiving content protects research and research integrity while minimizing researcher exposure to risk. Additionally, extremist content on social media and other platforms is often removed – arguably good for the public but challenging for future researchers. Internal databases of deleted content may be maintained by platforms but are rarely (if ever) accessible externally.

In an ideal operating environment, trusted researchers would be able to work with platforms and/or the government to maintain archives of harmful online content in a safe environment away from public exposure. Barring such a situation, the need to safely archive online content falls on individual researchers and their institutions. Open-Source Intelligence (OSINT) practices have vastly improved in recent years, which has led to many guides on carrying out online open-source research and safely archiving content. For example, the open-source investigative outfit Bellingcat provides a comprehensive toolkit (see Annex), while a quick search reveals multiple similar free and paid toolkits.

In general, archiving content is one of the best methods researchers have for analyzing and comparing information as the material is removed by platforms or by the user. This process should follow several guiding principles:

- Data in the form of images, screengrabs, videos, screen recordings, or documents should be downloaded from the digital source. OSINT recording tools can help in this process to record the entire investigative process automatically, while also allowing the analyst or researcher to streamline the archival process.
- If intended for wider use beyond the researcher or their immediate team, archived data should not contain any identifiable information such as the username or channel/page name, as this will prevent individuals outside of the research field from accessing the data's origin point. Despite extremists promoting hateful content, they still have a right to privacy, and removing identifiers prevents other

19 "Unicef Procedure for Ethical Standards in Research, Evaluation, Data Collection and Analysis," Unite for Children (UNICEF), April 1, 2015, https://www.unicef.org/media/54796/file#:~:text=In%20order%20to%20ensure%20the%20protection%20of%2C%20and,for%20ethical%20research%2C%20evaluation%20and%20data%20collection%20and.

individuals from gaining access to the data's location. If identifiers are not removed, it provides the opportunity for individuals from gaining access to the havens where extremism and hate thrive.

- Researchers should document for their own records the method used to locate the information for future research. Keeping detailed records of any archived content and any remarks a researcher may have upon first encountering the information online can help maintain the archive's organization. The value of information is not always obvious, so detailing the justification for archiving it is vital for recognizing its significance.
- Once the data has been collected and identifiable information is removed, the content should be archived. A secure location – safe online cloud services, encrypted Universal Serial Bus (USB) or external hard drives, or hard copies in a locked location may all be options. In general, if sensitive digital content should be encrypted prior to storing using tools like VeraCrypt, encrypted .7Z storage files with SHA256, or other alternative encryption techniques. The secure location should contain an organizational system to easily find content when needed for research purposes. Researchers should use a file-saving system that includes the date, location of data, and an identifiable characteristic of the content.

Considering exploitation

Extremism and terrorism research may have security or intelligence applications. Additionally, these subjects are inherently political. Political groups, states, and actors (both benign and malicious) may have an interest in promoting a position with regards to terrorism or extremism, having groups considered in a certain way, or exaggerating or minimizing a threat or social problem. Authoritarian regimes also use the threat of extremism or terrorism to change laws or erode civil liberties or use the language of 'terrorism' and 'extremism' to justify human rights violations. Any exploitation of analysis potentially opens a researcher or research institution (as well as the subjects of their research) to risk. Accordingly, when conducting analysis, it is important that researchers consider the ways their research may be misrepresented, taken out of context, or used to justify unethical actions or positions. While the risk of exploitation can not be eliminated, it should be considered at the inception and the publishing of research projects.

Trauma & Mental Health and Mitigating Vicarious / Secondary Trauma

Researchers are exposed to hateful content which can harm their mental health. Trauma occurs on a spectrum that varies by individual. Many researchers become desensitized to online violence because they see it so often. Such high levels of exposure to violence have been shown to often lead to emotional numbness and reduced response to real-world violence. It is often difficult to ask for help or express mental health issues, which is why institutions should be proactive in their promotion of mental health resources.

Handling emotionally laden and disturbing content on a consistent basis raises the likelihood of experiencing several negative effects. The risk of vicarious or secondary trauma and burnout among researchers in this field is particularly high. Vicarious trauma is defined as a "cognitive change through empathetic engagement with

trauma survivors."[20] This has been widely shown to impact interviewers and those who assist trauma survivors.[21] Secondary trauma stress is characterized as "a state of physical, emotional and mental exhaustion caused by long-term involvement in emotionally demanding situations."[22]

Viewing racism, hate, and violence can take a toll on a researcher's well-being, so it is vital for institutions to have resources in place. As discussed above, research institutions have a duty of care to protect staff from trauma and provide a space for discussions, opinions, and processing of the content seen online. Psychological stressors and traumatic events online can translate to physical health problems.[23] Institutions often provide mental health days for their employees outside of taking time off to allow time to rest. All research institutions and organizations can provide a variety of resources that can directly impact the mental health support an individual receives:

- Allowing researchers more mental health days to process graphic content they may have encountered that week. Mental health days should not count as paid time off or a sick day;
- Implementing hard limits on the amount of time a researcher can spend online; and,
- Provide weekly opportunities for researchers to openly discuss trauma, graphic material, and stressors with licensed professionals and/or among their peers.

There is research that suggests that accidental or unexpected exposure to harmful images can lead to greater psychological harm than that which is expected and undertaken with a clear sense of purpose. Accordingly, institutions should consider policies that prevent accidental exposure to this kind of content by other researchers or third parties. This could include the use of screen obscurers if an institution has an open office and the storing of potentially traumatic material on protected drives that other staff cannot access.

In addition to institutions implementing guidelines, researchers should find their own methods for coping with their work by choosing to engage in healthy behaviors that are non-work related. In a 2020 study, participants found peer support to be the most helpful for processing trauma because it provides immediate responses and colleagues can advise one another.[24] Additionally, it may be important for colleagues to engage in healthy dialogue regarding their research because some workplaces require employees to sign non-disclosure agreements that prevent them from talking about their research outside of the workplace.[25] Particularly compartmentalized or classified work that prevents colleagues from speaking to each other inside

20 Jason M. Newell & Gordon A. MacNeil, "Professional Burnout, Vicarious Trauma, Secondary Traumatic Stress, and Compassion Fatigue: A Review of Theoretical Terms, Risk Factors, and Preventative Methods for Clinicians and Researchers," *Best Practices in Mental Health: An International Journal* 6, no. 2 (July 14, 2010): 57–68, https://calio.org/wp-content/uploads/2020/04/Professional-burnout-vicarious-trauma-secondary-stress-and-compassion-fatigue-A-review-of-theoretical-terms-risk-factors-and-preventive-methods-for-clinicians-and-researchers.pdf.

21 Philip A. Sandick, "Speechless and Trauma: Why the International Criminal Court Needs a Public Interviewing Guide," *Northwestern Journal of International Human Rights* 11, no. 1 (2012): 105–125, https://scholarlycommons.law.northwestern.edu/njihr/vol11/iss1/4.

22 "Staff Well-Being and Mental Health in UNHCR," The United Nations High Commissioner for Refugees, 2016, https://www.unhcr.org/56e2dfa09.pdf; Andreas Seidler et al., "The Role of Psychosocial Working Conditions on Burnout and Its Core Component Emotional Exhaustion – a Systematic Review," *National Library of Medicine: National Center for Biotechnology Information* 9, no. 10 (March 14, 2014): 9–10, https://doi.org/10.1186/1745-6673-9-10.

23 "Mental Health Disorders and Stress Affect Working-Age Americans," Center for Disease Control, July 2018, https://www.cdc.gov/workplacehealthpromotion/tools-resources/pdfs/WHRC-Mental-Health-and-Stress-in-the-Workplac-Issue-Brief-H.pdf.

24 Lara Bakes-Denman, Yolanda Mansfield, & Tom Meehan, "Supporting Mental Health Staff Following Exposure to Occupational Violence – Staff Perceptions of 'Peer' Support," *International Journal of Mental Health Nursing* 30, no. 1 (February 2021): 158–166, https://doi.org/10.1111/inm.12767.

25 Cristina Criddle, "Facebook Moderator: 'Everyday was a nightmare'," BBC News, May 12, 2021, https://www.bbc.com/news/technology-57088382.

the workplace should be proactively addressed by making time for specific units or teams to debrief together. Other mitigation opportunities include:

- Encouraging researchers to engage in healthy activities that are non-work related to prevent work from "following" them home;
- Seeking out a licensed therapist to provide more individualized support and providing referrals to other licensed professionals if necessary; and,
- Ensuring that the research institution will provide mental health days (and proactively offer them) if the researcher needs to rest.

ANNEX: Researcher Safety Resources

Institutions

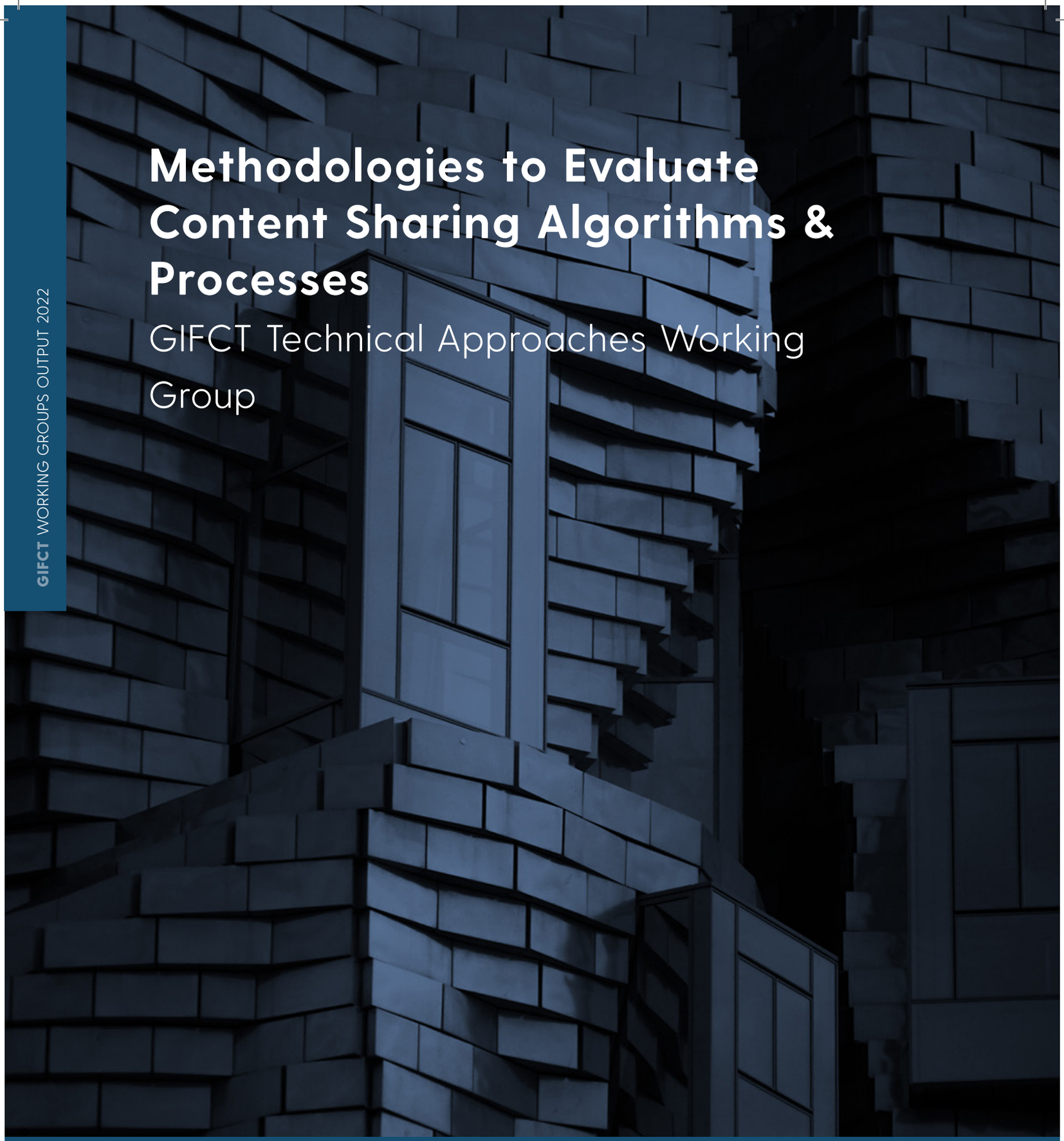
1. **Ethical Research Practices**
    a. Does the Institution Have a Plan for That? Researcher Safety and the Ethics of Institutional Responsibility
    b. The Development of the Framework for Research Ethics in Terrorism Studies (FRETS)
    c. Responsible Conduct of Research: Not Just for Researchers
    d. Ethics
    e. Do Researchers have an obligation to report dangerous actors?
    f. General Data Protection Regulation (GDPR)
2. **Protecting Researchers**
    a. Best Practices for Protecting Researchers and Research
    b. Being a woman on the internet is a nightmare. How can we fix it?
3. **Mental Health**
    a. The Center for Workplace Mental Health
    b. Supporting Mental Health in the Post-Pandemic Workplaces
4. **Archiving Content**
    a. Bellingcat's Online Investigation Toolkit
    b. Conifer

Individuals

1. **Safety**
    a. **General Online Safety**
        i. Researcher Welfare 1: Privacy and Security
        ii. Digital protection guides comparison
        iii. Digital protection guides - A survey of community resources
        iv. Crash Override Network: Resource Center
        v. Tactical Tech Data Detox
        vi. Surveillance Self-Defense: Tips, Tools and How-Tos for Safer Online Communications
        vii. Best VPN Service of 2022
        viii. NSA: Keeping Safe on Social Media
    b. **Online Harassment Protection**
        i. Speak Up & Stay Safe(r): A Guide to Protecting Yourself From Online Harassment
        ii. PUBLIC: PII Manager Opt-Out Template
        iii. Doxing: What it is and how to protect yourself
        iv. Report Illegal Content on the Internet: Europol
    c. **Other**
        i. UNESCO: Safety guide for journalists: a handbook for reporters in high-risk environments
2. **Mental Health**
    a. Researcher Welfare 2: Mental and Emotional Well-being and Self Care - VOX - Pol (voxpol.eu)
    b. The Body Keeps the Score: Brain, Mind, and Body in the Healing

   c. Headington Institute
      i. How Stressed You Are
      ii. You Are Showing Signs of Burnout
      iii. Understanding & Coping with Traumatic Stress
      iv. Preventing Burnout
      v. Understanding and Addressing Vicarious Trauma
3. The Center for Victims of Torture
      i. Self-Care Guide
      ii. Trauma Stewardship: An Everyday Guide to Caring for Self While Caring for Others
      iii. What About You? A Workbook for Those Who Care for Others
4. **Social Media**
   a. Introducing the Researcher Platform: Empowering independent research analyzing large-scale data from Meta
   b. Digital Void
   c. Center for Countering Digital Hate

Links to the Annex Resources can be found on the GIFCT Working Groups Webpage:
https://gifct.org/working-groups/



GIFCT WORKING GROUPS OUTPUT 2022

Methodologies to Evaluate Content Sharing Algorithms & Processes

GIFCT Technical Approaches Working Group

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,"[1] 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[2]. 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[3] – 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."[4] However, many questions remain about "borderline content"[5] 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.[6]

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.[7]

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 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

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-CAPI1-2021.pdf.

understanding of recommendation systems and user journeys."[8]

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."[9] 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[10] 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[11] 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, 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[12] also prohibit terrorist content. This is defined as material posted by or in support of

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.

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/.

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[13] 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 [14].

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

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/.

materials. YouTube describes borderline content as follows:

> **Content that comes close to – but doesn't quite cross the line of – violating our Community Guidelines.**[15]

While Meta has developed specific categories of borderline content, it too uses similar language in its community standards:[16]

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

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:

> **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.**[18]

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."[19]

15 "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/.

16 "Facebook Community Standards," Facebook, 2022, https://transparency.fb.com/policies/community-standards/.

17 "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/

18 "Content-Sharing Algorithms," GIFCT. https://gifct.org/wp-content/uploads/2021/07/GIFCT-CAPII-2021.pdf

19 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.

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 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 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."[20]

Yet the extent to which this is a true description of recommendation algorithms remains unknown. Although some

20 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.

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,[21] 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 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.

21 Chen et al., "Subscriptions and external links."

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[22] 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 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

22 "Inside TikTok's Algorithm: A WSJ Video Investigation," Wall Street Journal, July 21, 2021, https://www.wsj.com/articles/tiktok-algorithm-video-investigation-11626877477.

major platforms have already developed them for exactly this form of exploratory research.[23]

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 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.

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/.

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.[24] 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

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.

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.)[25] 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, trials are also conducted offline, for instance in so-called "multi armed bandit" simulations (see Bottou et al.).[26] 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).[27]

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"[28] or Rajthe et al.'s definition of "out-group language."[29]

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.

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.

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.[30] 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 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

30 Huszár, "Algorithmic amplification."

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.[31]

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.

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

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 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.[32] 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.[33] 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.

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.

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.

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.[34] 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 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%."[35] 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.

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.

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

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 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.[36] 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.[37] 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.[38]

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 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."[39]

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-andor-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.

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

To mitigate these risks, other mechanisms must be combined with anonymization and aggregation to improve privacy protection.[40] 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.[41] 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.[42]**

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.[43]

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

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-rasmussen/.

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/tumblrs-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.

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[44], misinformation, disinformation,[45] and conspiracy theories.[46] 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 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.[47] 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.

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/.

47 GIFCT Definitional Frameworks

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.[48] 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."[49] 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 interference with his privacy, family, home, or correspondence. Everyone has a right to the protection of the law against such interference or attacks."[50] 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"[51] include:

- Who gets access
- How data is used

48 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.

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

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.

- 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.[52] 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.[53] Conversely, as suggested in the second methodology, less direct access to data means that research 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).

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."

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).[54]

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.[55]

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 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).

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 LGBTIQA+ 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-lgbtiqa-community-and-role-eu-digital-services-act.

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."[56] 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 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

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

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.aaa1160.
- 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/ Indigenous 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 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 oppressed ethnic or racial 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 gender identity	Persons that fall within a spectrum of gender identities that are not exclusively masculine or feminine.
Homeless / 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[57]).
- **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.

GIFCT WORKING GROUPS OUTPUT 2022

Research Call for Proposals: Machine Translation

Technical Approaches Working Group

GIFCT
Global Internet Forum to Counter Terrorism

For development of a multilingual machine translation model capable of recognizing the nuanced use of language specific to a violent extremist context.

Background

The Global Internet Forum to Counter Terrorism (GIFCT) is a non-profit organization with a mission to prevent terrorists and violent extremists from exploiting digital platforms. Our vision is to build a world in which the technology sector marshals its collective creativity and capacity to render terrorists and violent extremists ineffective online. In every aspect of our work, we aim to be transparent, inclusive, and respectful of the universal human rights that terrorists and violent extremists seek to undermine.
Founded by Facebook, Microsoft, Twitter, and YouTube in 2017, GIFCT was established to foster technical collaboration among member technology companies, advance relevant research, and share knowledge. Since 2017, GIFCT's membership has expanded to include eighteen diverse digital platforms committed to cross-industry efforts to counter the spread of terrorist and violent extremist content online.

Three strategic objectives provide the focus for GIFCT to realize its vision:

1. Convene, engage, and provide thought leadership on the most important and complex issues at the intersection of terrorism and technology, demonstrating with concrete output that multistakeholderism can deliver genuine progress.
2. Create a global, diverse, and expansive community of GIFCT member companies reflective of the ever-evolving threat landscape.
3. Build the collective capacity and capability of the industry by offering cross-platform technology solutions, information sharing, and practical research for GIFCT members.

Content moderation is a challenging and resource-intensive task. The subject matter expertise required to accurately assess if a piece of content is terrorist or violent extremist material is rare and to achieve this level of understanding in multiple languages even more so. As critiques by groups such as EFF point out, "Automated technology doesn't work at scale; it can't read nuance in speech the way humans can, and for some languages it barely works at all. Over the years, we've seen the use of automation result in numerous wrongful takedowns. In short: automation is not a sufficient replacement for having a human in the loop."[1] As we design systems to support content moderation by skilled human reviewers, we should aim to ensure that they are provided the nuanced information they need in as accessible a format as possible.

Existing machine translation models can help to a certain extent. However, current multilingual models may not deeply model the subtleties of languages and language varieties to their full extent.[2] More concerning is how in many contexts, machine learning has been shown to contribute to (and potentially amplify) societal inequity,

1 Jillian York and Corynne McSherry, "Automated Moderation Must be Temporary, Transparent and Easily Appealable," Electronic Frontier Foundation, April 2, 2020, https://www.eff.org/deeplinks/2020/04/automated-moderation-must-be-temporary-transparent-and-easily-appealable.

2 Zihan Wang et al., "Extending Multilingual BERT to Low-Resource Languages," Findings of the Association for Computational Linguistics: EMNLP 2020, (November, 2020): 2649–2656, https://aclanthology.org/2020.findings-emnlp.240/.

furthering the unjust treatment of people who have been historically discriminated against.[3] While inequity is not an inevitable consequence of these models, it is essential to identify such potential effects through proactive and reactive means.

Aims

GIFCT seeks the development of a multilingual machine translation system that is capable of recognizing the nuanced use of language specific to a violent extremist context, enabling subject matter experts to apply it toward moderating content efficiently in multiple languages. We are also conscious of the human rights and ethical implications of applying such technologies and seek to apply a human rights-based approach to the development, evaluation, and application of any solutions developed.

Requirements:

- A solution that is capable of recognizing the nuanced use of language specific to a violent extremist context.
- The model will be capable of translating text from multiple different languages into a target language (English).
- The model can be executed using a standard machine learning framework such as PyTorch or TensorFlow.
- The model and the process of building and training the model will be shown to provide sufficient data protection to protect user privacy in line with GDPR and other regulations.
- The vendor will be shown to have taken reasonable steps to identify and address potential issues of bias in the model and in the process of building and training the model.
- To the extent the model includes or is integrated with any third party intellectual property, such IP will preferably be licensed as open-source (e.g., https://opensource.org/licenses/), or alternatively (and only after consultation with GIFCT), can be made available in perpetuity with a fully paid license in favor of GIFCT and associated entities and persons for use in preventing terrorist or violent extremist content.
- Vendors must not have any conflict of interest with GIFCT or GIFCT staff.

Evaluation

- Performance of the model across a broad range of languages in line with the XTREME benchmark
- Performance of the model to capture the cultural specificity of different violent extremist groups' use of language
- Alignment with GIFCT's Mission and Values
- Sensitivity to Human Rights and Ethical issues that may arise

Deadlines and Format

3 Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models," Center for Research on Foundation Models (CRFM), August 16, 2021, https://fsi.stanford.edu/publication/opportunities-and-risks-foundation-models.

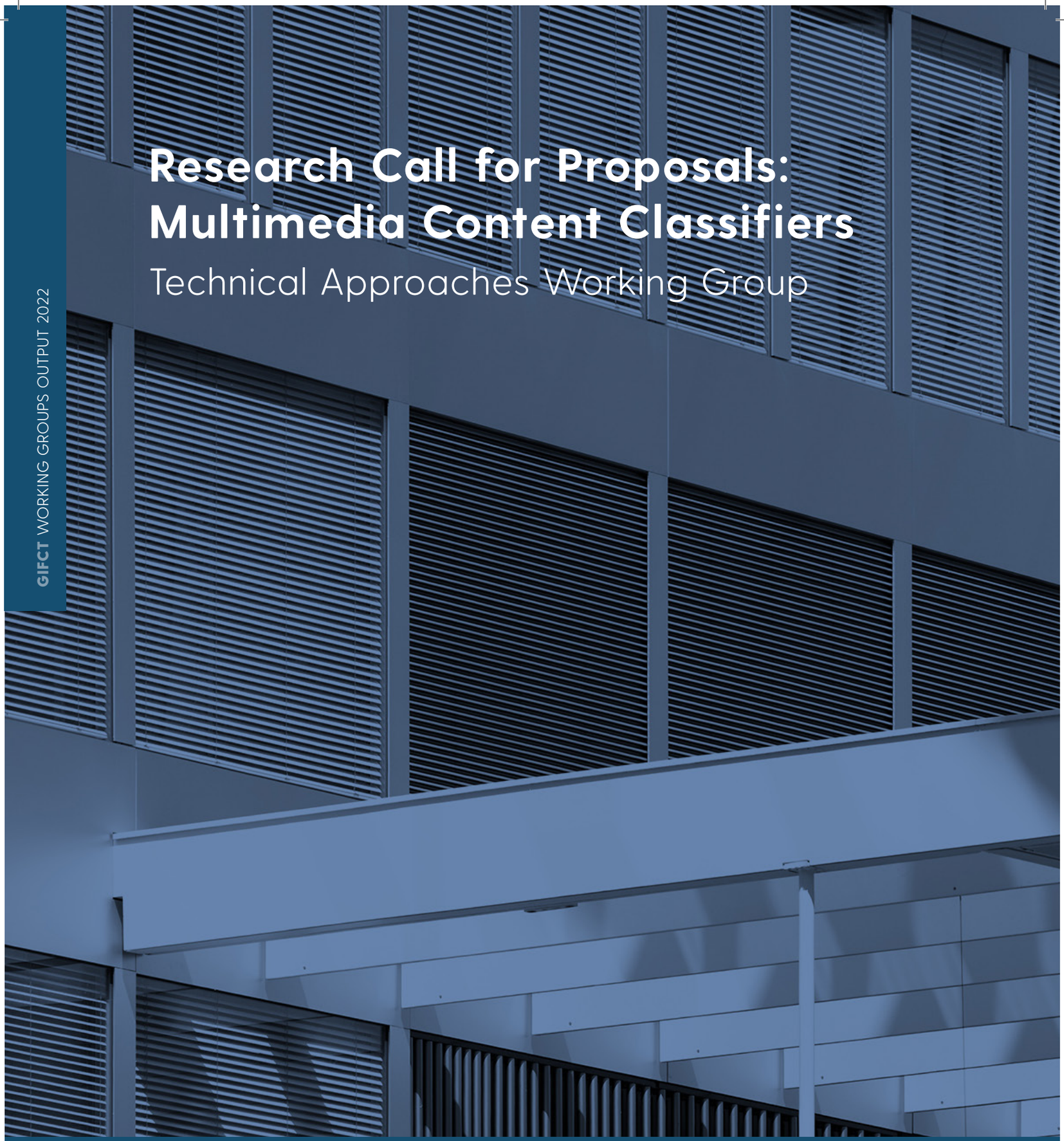
GIFCT aims to begin the research project in July 2022 for completion before the end of the calendar year. Proposals or other inquiries should be submitted to tech@gifct.org with the email subject line “Machine Translation Proposal" by Friday, June 24, 2022. Proposals should include the following:

- A brief overview of your proposed approach to this project
- An estimated timeline for delivery
- An estimated budget or costs
- A brief overview of the organization or team that would deliver the project

GIFCT WORKING GROUPS OUTPUT 2022

Research Call for Proposals: Multimedia Content Classifiers

Technical Approaches Working Group



GIFCT
Global Internet Forum to Counter Terrorism

For development of a system to classify multimedia content as terrorist or violent extremist content.

Background

The Global Internet Forum to Counter Terrorism (GIFCT) is a non-profit organization with a mission to prevent terrorists and violent extremists from exploiting digital platforms. Our vision is to build a world in which the technology sector marshals its collective creativity and capacity to render terrorists and violent extremists ineffective online. In every aspect of our work, we aim to be transparent, inclusive, and respectful of the fundamental and universal human rights that terrorists and violent extremists seek to undermine.

Founded by Facebook, Microsoft, Twitter, and YouTube in 2017, GIFCT was established to foster technical collaboration among member technology companies, advance relevant research, and share knowledge with all our member companies. Since 2017, GIFCT's membership has expanded beyond the founding companies to include eighteen diverse digital platforms committed to cross-industry efforts to counter the spread of terrorist and violent extremist content online.

Three strategic objectives provide the focus for GIFCT to realize its vision:

1. Be a leading organization to convene, engage, and provide thought leadership on the most important and complex issues at the intersection of terrorism and technology, demonstrating with concrete output that multistakeholderism can deliver genuine progress.
2. Create a global, diverse, and expansive community of GIFCT member companies reflective of the ever-evolving threat landscape.
3. Build the collective capacity and capability of the industry by offering cross-platform technology solutions, information sharing, and practical research for GIFCT members.

Content moderators need to make decisions about whether specific content violates the content policies of social media platforms. Given the large volume and breadth of content, it is important to be able to prioritize specific content for human moderation.

As we design systems to support content moderation by skilled human reviewers, we should aim to ensure that they are provided the nuanced information that they need in as accessible a format as possible.

In many contexts, machine learning has been shown to contribute to, and potentially amplify, societal inequity, furthering the unjust treatment of people who have been historically discriminated against[1]. While inequity is not an inevitable consequence of these models, it is essential to identify such potential effects through proactive and reactive means.

1 Bommasani, R. (2021, August 16). On the Opportunities and Risks of Foundation Models. THE FREEMAN SPOGLI INSTITUTE FOR INTERNATIONAL STUDIES. https://fsi.stanford.edu/publication/opportunities-and-risks-foundation-models.

Aims

GIFCT is seeking proposals for the development of a system to classify multimedia content as conforming to some definition of terrorist and violent extremist content in a way that is contextualized and explainable, and provides some degree of confidence or probability to the user (hereinafter, the "Solution"). The Solution is intended to be used to inform human content moderators decisions about terrorist and violent extremist content and help prioritize their reviews.

Requirements

- Given some definition of violent extremism, the Solution should be able to classify content as belonging to that definition or not, and to what probability or confidence that it is part of that definition.
- The Solution classifies content as being violent extremist or terrorist content based on a broad understanding from experts in the field.
- The Solution will provide sufficient context and explanation to the user that can help determine why the classification decision was made.
- The Solution will be able to classify at least one content type but may also be multi-modal and consider content types such as audio, video, images, text.
- The Solution can be executed using a standard machine learning framework such as PyTorch or TensorFlow.
- The Solution can be further fine-tuned by users through training it on additional content.
- The Solution and the process of building and training the Solution will be shown to provide sufficient data protection to protect user privacy in line with GDPR and other regulations.
- The vendor will be shown to have taken reasonable steps to identify and address potential issues of bias in the Solution and in the process of building and training the Solution.
- To the extent the Solution includes or is integrated with any third-party intellectual property, such IP will preferably be licensed under an open source license (such as those listed here: https://opensource.org/licenses/), or alternatively, and only after consultation with GIFCT, can be made available with a perpetual, fully paid-up license in favor of GIFCT and associated entities and persons for use in preventing terrorist or violent extremist content.

Evaluation

- Performance of the Solution across a broad range of content types using metrics such as precision and recall.
- Performance of the Solution to capture a broad range of violent extremist groups and ideologies.
- Alignment with GIFCT's Mission.
- Alignment with GIFCT's Values.
- Sensitivity to Human Rights and Ethical issues that may arise.

Deadlines and Format

GIFCT aims to begin the research project in September 2022 for completion by July 2023.

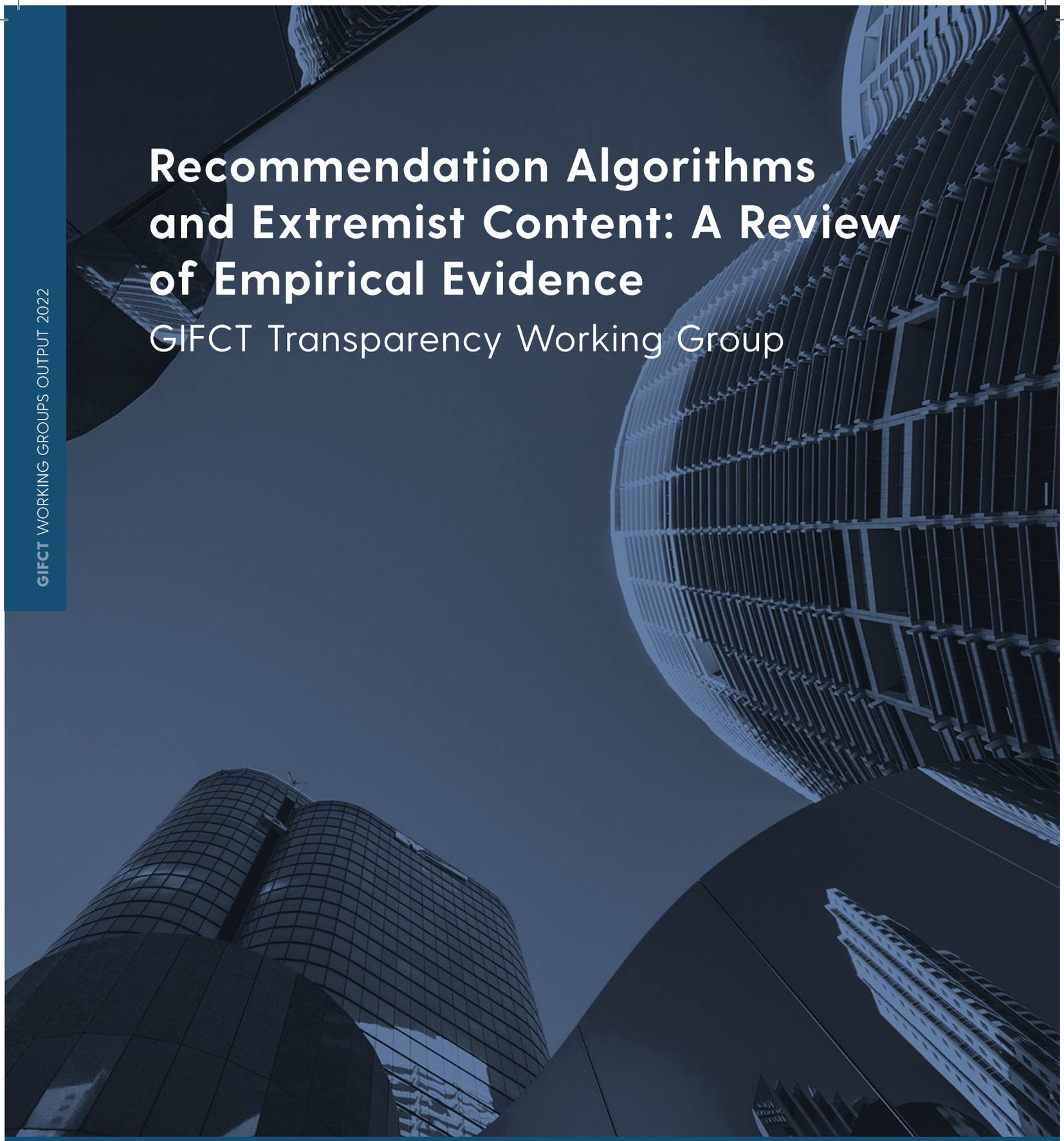
Proposals or other inquiries should be submitted to tech@gifct.org with the email subject "Classifier Proposal" by Friday, June 24, 2022 and proposals should include:

- A brief overview of your proposed approach to this project
- An estimated timeline for delivery
- An estimated budget or costs
- A brief overview of the organization or team that would deliver the project

GIFCT WORKING GROUPS OUTPUT 2022

Recommendation Algorithms and Extremist Content: A Review of Empirical Evidence

GIFCT Transparency Working Group



GIFCT
Global Internet Forum to Counter Terrorism

Dr. Joe Whittaker
Swansea University

Introduction

This paper reviews the existing empirical studies on the role of social media recommendation algorithms and potential links to extremist content. It seeks to provide transparency for future researchers by taking stock of the present empirical knowledge base, noting the types of data and methods that are utilized and charting gaps in the research. In doing so, the paper sheds light on definitional issues, replicability, agency, causality, and the limitations of present approaches. It also offers a window into how social media companies have adapted their practices over the past decade by surveying public statements about their policies. In total, 15 studies were identified for review. The appendix contains a table outlining each study's methods and findings.

Out of the review, a nuanced picture emerges of research into the role of recommendation systems and extremist content. There is a heavy emphasis on studying YouTube, a focus on the far-right, as well as English-language content. Moreover, many of the studies collect and analyze data in the mid-2010s, which importantly was before many platforms began to downrank or remove borderline content from recommendations. Although there is a wide array of methods utilized to investigate this topic, all but two papers rely on an external "black box" methodology in which researchers cannot manipulate platforms' recommendation systems. Similarly, there are only three studies in total which utilize a control group. More often, researchers access a platform's Application Programming Interface (API) to establish content that could potentially be recommended, which cannot account for personalization. There are also substantial differences in how "extremism" or related concepts are deployed in coding systems, which is a challenge for meta-reviews such as this. Most of the studies show that platforms can recommend extremist content, although there are several important caveats, such as the importance of pre-existing user beliefs and related variables. This review concludes by contextualizing the findings in the wider academic and policy debates, as well as offering a set of recommendations moving forward.

Before beginning, it is worthwhile to be clear about the topic under investigation. The Global Internet Forum to Counter Terrorism's (GIFCT) Content-Sharing Algorithms, Processes, and Positive Interventions Working Group highlights three categories of social media algorithms that could be exploited by violent extremists:

i. **Search algorithms** – such as autocompleting a keyword;
ii. **Recommendation algorithms** – which curate content that a user may be interested in; and
iii. **Ad-tech algorithms** – that target users based on demographics and behavior to optimize advertising.[1]

This review is focused specifically on recommendation algorithms, which Ricci, Rokach, and Shapira define as software tools and techniques which provide users with suggestions for items a user may wish to utilize.[2] An important distinction is whether the content is pre-selected or user-selected.[3] For example, YouTube's "Recommended for you" algorithm offers algorithmically driven suggestions that the user has the option to select or ignore; on the other hand, platforms with news feeds or timelines do not necessarily offer this choice,

1 "Content-Sharing Algorithms, Processes, and Positive Interventions Working Group," Global Internet Forum to Counter Terrorism, 2021, https://gifct.org/wp-content/uploads/2021/07/GIFCT-CAPII-2021.pdf.

2 Frencesco Ricci, Lior Rokach, and Bracha Shapira, Recommender Systems Handbook, (New York: Springer, 2011).

3 Frederik J. Zuiderveen Borgesius et al., "Should we worry about filter bubbles?" Internet Policy Review 5, no. 1 (March 31, 2016). https://doi.org/10.14763/2016.1.401.

with content appearing before a user has selected it.[4] This review follows the lead of the above-mentioned GIFCT report in including both pre-selected and user-selected recommendation systems – i.e. inclusive of timelines, news feeds and recommended videos, etc.[5]

The scope of this review spans studies that analyze extremist content. However, "extremist" is an essentially contested concept,[6] as are related words such as "terrorist" and "radical," which are often used interchangeably, and which have led to considerable conceptual ambiguity.[7] Depending on one's conceptualization, extremist content on social media may include terrorist propaganda, materials that directly advocate violence, incitement of hatred, and/or non-illegal yet potentially harmful content that may "other" certain out-groups. To be included in this review, the authors of a study must identify the content as extremist, radical, or terrorist, or refer to specific ideologies that are widely considered to be extreme (such as the far-right or jihadism). While this is an imperfect criterion, it is worthwhile to be inclusive, and then utilize this review to analyze the decision-making of the researchers and coding systems in the corpus of literature to determine what kind of content is deemed extreme.

Given this scope, this review does not include research into the wider field of political discourse and polarization. Therefore, studies such as Bakshy, Messing, and Adamic's research into Facebook's News Feed or Cho and colleagues' laboratory experiment of YouTube's recommendation system are omitted.[8] Although there is a link between polarization and extremism,[9] as well as concerns over the normalization of far-right narratives into mainstream politics,[10] this review is focused specifically on content that has been explicitly identified as extreme. Similarly, there is a growing empirical literature on recommendation systems and disinformation or misinformation.[11] While there is an overlap between disinformation and extremism,[12] only studies that are explicitly related to extremism have been included. One study was excluded despite making specific reference to "extreme content" in the title as the study did not actually analyze extremism but instead focused on

4 Joe Whittaker et al., "Recommender Systems and the Amplification of Extremist Content," Internet Policy Review, 10, no. 2 (June 30, 2021), https://policyreview.info/articles/analysis/recommender-systems-and-amplification-extremist-content.

5 "Content-Sharing Algorithms," GIFCT.

6 Walter Bryce Gallie, "Essentially Contested Concepts," Proceedings of the Aristotelian Society, 56, (1955): 167–198.

7 For example, see Randy Borum, "The Etiology of Radicalization," In The Handbook of the Criminology of Terrorism, eds. G. LaFree and J. D. Freilich (Chichester: John Wiley and Sons, 2017): 17–32; Bart Schuurman and Max Taylor, "Reconsidering Radicalization: Fanaticism and the Link Between Ideas and Violence," Perspectives on Terrorism, 12, no. 1 (2018): 3–22.

8 Eytan Bakshy, Solomon Messing and Lada A. Adamic, "Exposure to Ideologically Diverse News and Opinion on Facebook," Science Express, (May, 2015), 1–5, doi: 10.1111/j.1460-2466.2008.00410.x.; Jaeho Cho et al., "Do Search Algorithms Endanger Democracy? An Experimental Investigation of Algorithm Effects on Political Polarization," Journal of Broadcasting and Electronic Media 64, no. 2 (2020): 150–172.

9 Cass R. Sunstein, "The law of group polarization," The Journal of Political Philosophy, 10, no. 2, (2002): 175–195; Bertjan Doosje et al., "Terrorism, Radicalization and De-radicalization," Current Opinion in Psychology 11 (2016): 79–84.

10 Aurelien Mondon and Aaron Winter, "Articulations of Islamophobia: From the extreme to the mainstream?," Ethnic and Racial Studies, 40, no. 13 (2017): 2151–2179.

11 For example, see Eslam Hussein, Prerna Juneja, and Tanushree Mitra, "Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube," Proceedings of the ACM on Human-Computer Interaction, 4 (2020), doi: 10.1145/3392854; Konstantinos Papadamou et al., "'It is just a flu': Assessing the Effect of Watch History on YouTube's Pseudoscientific Video Recommendations," (2020), http://arxiv.org/abs/2010.11638; Jonas Kaiser, Adrian Rauchfleisch, and Yasodara Córdova, "Fighting Zika With Honey: An Analysis of YouTube's Video Recommendations on Brazilian YouTube," International Journal of Communication 15, no. 108 (2021): 1244–1262.

12 Jamie Bartlett and Carl Miller, "The Edge of Violence: Towards Telling the Difference Between Violent and Non-Violent Radicalization," Terrorism and Political Violence, 24, no. 1 (2012), 1–21; Jacob Davey and Julia Ebner, "The Great Replacement: The Violent Consequences of Mainstreamed Extremism," Institute for Strategic Dialogue (2019), https://www.isdglobal.org/wp-content/uploads/2019/07/The-Great-Replacement-The-Violent-Consequences-of-Mainstreamed-Extremism-by-ISD.pdf.

"contextually inappropriate" recommendations such as satire and sexually suggestive material being found by children.[13]

Data

YouTube is the most popular platform for empirical research within the existing literature; nine of the studies analyzed whether its recommendation system promotes extreme content. This is followed by Twitter, whose recommendations or timeline are analyzed in three studies. Two studies explore Reddit, either looking at the up and downvoting system or its "Best" timeline, with one study each looking at Facebook and Gab. Finally, one study is interview-based and does not explicitly focus on one platform, although the findings mention YouTube's recommendation system. YouTube's dominance in the literature is highlighted by Whittaker and colleagues, who note that it has a researcher-friendly API compared to other social media platforms; they suggest that the field may be driven by research convenience rather than necessarily following the trail of extreme content.[14]

The studies are weighed towards researching far-right content. This is counter to the general pattern in the wider field of terrorism and extremism studies, which has often been noted as being primarily focused on jihadism.[15] Six of the studies examine far-right content exclusively, while three others study the far-right and another ideology (such as jihadism, male supremacism, or the far-left). Five focus exclusively on jihadism or Islamism, while one analyses the incel community. The emphasis on the far-right may be related to the relative success with which social media platforms have been able to identify and remove jihadist content and the difficulties in applying the same lessons to far-right content.[16] In other words, researchers may be choosing to use far-right data because they have not yet been removed.

Perhaps owing in part to the focus on the far-right, there is also an English-speaking and Western focus to the data in existing studies. In seven cases, the "seed" accounts that are utilized to begin data collection identify English-speaking accounts either exclusively or predominantly. German content is also utilized; twice in studies with one English and one German-speaking dataset, and once exclusively in German. Two studies utilize Arabic-language content, and three studies had a mixed set of languages which include English, Japanese, French, Spanish, German, Turkish, Arabic, and Mandarin.

Research Within a Changing Social Media Landscape

A relevant factor within the corpus of academic literature is when the data for each individual study are collected. Two studies collected data in 2013,[17] Murthy's research uses a dataset from 2016,[18] with two more in

13 Christian Stöcker and Mike Preuss, "Riding the Wave of Misclassification: How we end up with extreme YouTube content," Lecture Notes in Computer Science (July 10, 2020): 359–375.

14 Whittaker et al., "Recommender Systems."

15 Maura Conway, "Determining the Role of the Internet in Violent Extremism and Terrorism: Six Suggestions for Progressing Research," Studies in Conflict and Terrorism, 40, no. 1 (2017): 77–98; Bart Schuurman, "Topics in Terrorism Research: Reviewing trends and gaps, 2007–2016," Critical Studies on Terrorism, 12, no. 3 (2019): 463–480.

16 Maura Conway, "Routing the Extreme Right: Challenges for Social Media Platforms," RUSI Journal 165, no. 1 (2020): 108–113.

17 Derek O'Callaghan et al., "Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems," Social Science Computer Review 33, no. 4 (2015), 459–478; J. M. Berger, "Zero Degrees of al Qaeda," Foreign Policy, August 14, 2013, http://foreignpolicy.com/2013/08/14/zero-degrees-of-al-qaeda/.

18 Dhiraj Murthy, "Evaluating Platform Accountability: Terrorist Content on YouTube," American Behavioral Scientist, 65, no. 6 (2021): 800–824.

2017[19] and one study spanning October 2017 to March 2018.[20] Ledwich and Zaitsev's study does not explicitly state the date of data collection, but it can be inferred that it likely took place in 2019.[21] Four further studies use data from that year,[22] with one study doing so in 2020.[23] Two take longitudinal approaches that span multiple years: Hosseinmardi et al. collect viewing behaviors from January 2016 to December 2019 and Huszár et al. access timeline data for users from June 2016 to June 2020.[24] The article by Baugut and Neumann does not specify dates for reasons of anonymity, but the interviews were conducted a relatively short time before publication.[25] As this study involved participants reflecting on their propaganda use in the past, it is difficult to assign a general timeframe.

Understanding when these studies collected data is important because the landscape of social media regulation has changed dramatically in the past decade. Academics point to 2016 as a turning point in which platforms began to take a more proactive approach toward removing content and disrupting accounts sympathetic to terrorism.[26] Prior to this, sites often framed themselves as allowing free speech to run its course on their sites. In 2012, Twitter's then-General Manager Tony Wang noted that: "We remain neutral as to the content because our general counsel and CEO like to say that we are the free speech wing of the free speech party."[27] Similarly, YouTube's decision to remove non-violent videos of Anwar al-Awlaki inciting violence in 2017 represented a key policy change, which saw a more proactive approach towards terrorist content.[28] This does not mean that terrorist content was not removed prior to 2016, but that platforms have become more proactive and as a result more sophisticated at detecting and removing content. In short, early studies in the cohort may draw from data that would not be recommended today. Murthy makes this point explicitly, noting that his dataset (derived from 2016) pre-dates the formation of GIFCT, which signaled increased efforts to prevent terrorists and violent extremists from exploiting digital platforms.[29]

19 Josephine B. Schmitt et al., "Counter-messages as prevention or promotion of extremism?! The potential role of YouTube," Journal of Communication 68, no. 4 (2018): 758–779; Tiana Gaudette, Ryan Scrivens, and Garth Davies, "Upvoting Extremism: Collective identity formation and the extreme right on Reddit," New Media and Society (September, 2020), doi: 10.1177/1461444820958123.

20 Gregory Waters and Robert Postings, "Spiders of the Caliphate: Mapping the Islamic State's Global Support Network on Facebook," Counter-Extremism Project (May, 2018).

21 Mark Ledwich and Anna Zaitsev, "Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization," Eprint arXiv:1912.11211, (2019).

22 Manoel H. Ribeiro et al., "Auditing Radicalization Pathways on YouTube," Woodstock '18: ACM Symposium on Neural Gaze Detection (2019) http://arxiv.org/abs/1908.08313; Michael Wolfowicz, David Weisburd and Badi Hasisi, "Examining the interactive effects of the filter bubble and the echo chamber on radicalization," Journal of Experimental Criminology (2021), doi:10.1007/s11292-021-09471-0.1; Whittaker et al., "Recommender Systems,"; Kostantinos Papadamou et al., "How over is it?" Understanding the Incel Community on YouTube," Proceedings of the ACM on Human-Computer Interaction 5 (2021), https://arxiv.org/abs/2001.08293.

23 Annie Y. Chen et al., "Exposure to Alternative and Extremist Content on YouTube," Anti-Defamation League, https://www.adl.org/resources/reports/exposure-to-alternative-extremist-content-on-youtube.

24 Homa Hosseinmardi et al., "Evaluating the scale, growth, and origins of right-wing echo chambers on YouTube" (2020), arXiv; Ferenc Huszár et al., "Algorithmic Amplification of Politics on Twitter," Proceedings for the National Academy of Sciences of the United States of America, 119, no. 1 (2022).

25 Information provided by Dr. Philip Baugut by email; see Philip Baugut and Katharina Neumann, "Online propaganda use during Islamist radicalization," Information Communication and Society, 23, no. 11 (2020): 1570–1592.

26 J. M. Berger and Heather Perez, "The Islamic State's diminishing returns on Twitter: How suspensions are limiting the social networks of English-speaking ISIS supporters," George Washington University: Program on Extremism, (February 2016); Maura Conway, "Violent Extremism and Terrorism Online in 2016: The Year in Review," Vox Pol (2016).

27 Josh Halliday, "Twitter's Tony Wang: 'We are the free speech wing of the free speech party,'" The Guardian March 22, 2012, https://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech.

28 Scott Shane, "In 'Watershed Moment,' YouTube Blocks Extremist Cleric's Message," New York Times, November 27, 2017, https://www.nytimes.com/2017/11/12/us/politics/youtube-terrorism-anwar-al-awlaki.html.

29 Murthy, "Evaluating Platform Accountability."

While content removal is an important factor – extremist material cannot be recommended if it is not available on the platform – given the scope of this review, it is worthwhile to consider how platforms have changed their approach to recommendation systems when faced with potentially problematic content that does not clearly violate some aspects of platforms' rules or terms of service. Below, an overview of policy for each of the five different platforms under study (YouTube, Facebook, Twitter, Reddit, and Gab) within the corpus of literature is presented to ascertain whether (and if so how) they changed their recommendations in the face of extreme content.

The first platform to publicly announce an alteration to recommendations was Reddit, which introduced a policy of "quarantining" subreddits in 2015. When the platform applies these measures, the subreddit is only viewable to those who explicitly opt-in.[30] This approach is taken to "prevent its content from being accidentally viewed by those who do not knowingly wish to do so, or viewed without appropriate context."[31] Quarantined subreddits do not appear in non-subscription-based feeds (such as Reddit's "Popular" feed) and are not included in search or recommendations. This is relevant for both studies on Reddit in this corpus: Gaudette and colleagues collect data from r/The_Donald in 2017, which was subsequently quarantined in 2019 before eventually being banned.[32] On the other hand, Whittaker et al.'s study found that Reddit's "Best" timeline did not recommend extreme content, which could be a result of having removed problematic content due to quarantining.[33]

YouTube takes a four-pronged approach to content moderation: Removing problematic and violative content from the platform; Raising up authoritative voices; Rewarding trusted creators; and Reducing the recommendations of borderline content.[34] The tactic of Reducing was first articulated in 2017 as a counter-terrorism policy, noting that the platform would take a "tougher stance on videos that do not clearly violate policies but may be inflammatory or supremacist" by removing content from being recommended.[35] According to YouTube, this step reduced views of such videos by an average of 80%.[36] This tactic was expanded to misinformation and conspiracy theories in early 2019 (YouTube 2019c),[37] which saw a drop of 70% in views of this content.[38] In mid-2019, the platform announced that they were expanding this policy by including authoritative voices into potential recommendations when an individual is watching borderline content.[39] YouTube also reward

30 Reddit, Content Policy Update, (2015), https://www.reddit.com/r/announcements/comments/3fx2au/content_policy_update/.

31 Reddit, Quarantined Subreddits, (2021), https://www.reddithelp.com/hc/en-us/articles/360043069012.

32 Gaudette, Scrivens, and Davies, "Upvoting Extremism."

33 Whittaker et al., "Recommender Systems."

34 "The Four Rs of Responsibility, Part 1: Removing harmful content," YouTube (September 3, 2019) https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove/.

35 Kent Walker, "Four steps we're taking today to fight terrorism online," Google, June 18, 2017, https://blog.google/around-the-globe/google-europe/four-steps-were-taking-today-fight-online-terror/.

36 "Our Ongoing Work to Tackle Hate," YouTube, June 5, 2019, https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate/.

37 "Continuing Our Work to Improve Recommendations on YouTube," YouTube, January 25, 2019, https://blog.youtube/news-and-events/continuing-our-work-to-improve/.

38 "Managing Harmful Conspiracy Theories on YouTube," YouTube, October 15, 2020, https://blog.youtube/news-and-events/harmful-conspiracy-theories-youtube/.

39 "Our Ongoing Work."

their trusted creators financially through their monetization program.[40] Although YouTube's policy has been updated several times, several studies collected data before the initial changes in 2017 (for example, the studies by O'Callaghan et al., Schmitt et al., and Murthy) while others did so after the first policy but before more recent updates, such as the public articulation of the 4Rs framework, including the research by Ledwich & Zaitsev, Ribeiro et al., and Whittaker et al., with the study by Chen and colleagues taking place after all of the updates outlined above. The data collected in the longitudinal studies (i.e. Hosseinmardi et al. and Papadamou et al.) reflects several changes in recommendation policy.

Facebook takes an approach similar to YouTube's, demarcating their moderation policy into three prongs: Removing violative content, Reducing misleading content via ranking and Informing users with additional context. This policy was outlined in 2018, mostly framed around sensationalist material and clickbait, noting that problematic content that does not violate policies can still be harmful to users, and when identified was downranked in the platform's News Feed.[41] However, Facebook's policy has since been updated to explicitly include content that may incite hatred, particularly in countries at risk of conflict.[42] Facebook also operates a Dangerous Individuals and Organizations policy, which seeks to restrict recommending movements that may be tied to violence but do not meet the criteria to be banned. The pages of groups designated as such by this policy are not eligible to be recommended and are downranked in the News Feed, as well as not suggested in the search function.[43] Only one study in the corpus studies Facebook and its findings relate to friend recommendations, so it is unclear whether its data is implicated by any of these policies.[44]

Twitter updated its Hateful Conduct Policy in 2019 to include several different enforcement options for dealing with hate speech.[45] This included existing consequences such as account suspension and the removal of tweets, but also the ability to downrank tweets within replies, making tweets ineligible for amplification in "Top Search" and/or on timelines for users that do not follow the author and excluding tweets and accounts in email or in-product recommendations.[46] Each of the three studies in this corpus began data collection before this policy change, although one study is longitudinal and the new policy was implemented during its time frame.[47]

Finally, Gab does not appear to have a policy that removes content from their recommendations (such as their "Popular" timeline). Gab claims that they use the First Amendment as their guiding principle and "make the best effort to ensure that all content moderation decisions and enforcement... does not punish users for exercising

40 Neal Mohan, "Responsibility is good for business and for the economy," YouTube (Blog), August 23, 2021, https://blog.youtube/inside-youtube/responsibility-good-business-and-creator-economy/

41 Tessa Lyons, "The Three-Part Recipe for Cleaning up Your News Feed," Facebook Newsroom, May 22, 2018, https://about.fb.com/news/2018/05/inside-feed-reduce-remove-inform/; Mark Zuckerberg, "A Blueprint for Content Governance and Enforcement," Facebook, November 15, 2018, https://www.facebook.com/notes/751449002072082/.

42 Samidh Chakrabarti and Rosa Birch, "Understanding Social Media and Conflict," Facebook Newsroom, June 20, 2019, https://about.fb.com/news/2019/06/social-media-and-conflict/.

43 "An Update to How We Address Movements and Organizations Tied to Violence," Facebook Newsroom, August 19, 2020, https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/.

44 Waters and Postings, "Spiders of the Caliphate."

45 Twitter, "Updating our Rules Against Hateful Conduct," July 9, 2019, https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.

46 Twitter, "Hateful Conduct Policy," https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

47 Huszár et al., "Algorithmic Amplification."

their... right to speak freely".[48] They state that they reserve the right to remove content or ban accounts that they feel violate the First Amendment's protection, but do not mention any kind of algorithmic downranking.

Methods

Research Objective

The role of recommendation systems is not the primary research objective or dependent variable in every study in the corpus. It is the main focus in seven,[49] for three it was just one of several research questions,[50] while for two, it is tested alongside another variable, such as proxies for an "echo chamber" effect.[51] The research of Huszár and colleagues is primarily concerned with recommendation algorithms, but the proliferation of far-right and far-left accounts is only one of several research questions.[52] Some studies had other goals, which led to findings that offer a perspective on algorithms. Waters and Postings conduct an analysis of online supporters and had incidental findings that relate to Facebook's recommendation system.[53] Similarly, Baugut and Neumann seek to understand the media diet of radical Islamists, who in turn self-report the importance of social media algorithms.[54]

Internal Versus External Access

A key distinction is how researchers accessed the respective datasets. In their report on responsible AI for social media, Knott and colleagues dichotomize between "Internal" access – i.e. studies in which social media companies work with researchers and provide private privileged data access – and "External" access – which uses publicly available datasets.[55] In this review, fourteen of the studies accessed external data, while only one relied on internal access: Huszár et al.'s study involved internal Twitter data and a research team comprising platform staff and academics.[56] The vast majority of studies in this corpus rely on "black box" testing in which researchers input data and receive outputs without having an understanding of how the underlying algorithms makes decisions.[57] Knott et al. argue that this discrepancy is a key limitation of the academic body of knowledge; external studies do not test causal hypotheses about the effects of recommender systems. Some of these types of studies, they argue, can manipulate users (either real or automated), but none of them can manipulate the platforms' recommendation system to observe its effect on users.[58] This has important ramifications: it is difficult to achieve algorithmic transparency given the "black box" nature of platforms' recommendation systems, which are closely guarded trade

48 Gab, Website Terms of Service, April 10, 2020, https://gab.com/about/tos.

49 Berger, "Zero Degrees"; O'Callaghan et al., "Down the (White) Rabbit Hole"; Schmitt et al., "Counter-messages as prevention"; Ledwich and Zaitsev, "Algorithmic Extremism"; Gaudette, Scrivens, and Davies, "Upvoting Extremism"; Whittaker et al., "Recommender Systems"; Murthy, "Evaluating Platform Accountability."

50 Ribeiro et al., "Auditing Radicalization Pathways"; Papadamou et al., "How over is it?"; Chen et al., "Exposure to Alternative & Extremist Content."

51 Hosseinmardi et al., "Evaluating"; Wolfowicz, Weisburd, and Hasisi, "Examining the interactive effects."

52 Huszár et al., "Algorithmic Amplification."

53 Waters and Postings, "Spiders of the Caliphate."

54 Baugut and Neumann, "Online propaganda use."

55 Alistair Knott et al., "Responsible AI for Social Media," The Global Partnership on Artificial Intelligence, (2021), https://gpai.ai/projects/responsible-ai/social-media-governance/responsible-ai-for-social-media-governance.pdf.

56 Huszár et al., "Algorithmic Amplification."

57 "Content-Sharing Algorithms," GIFCT.

58 Alistair Knott et al., "Responsible AI for Social Media."

secrets as exposing their inner workings may lead to a competitive disadvantage[59] as well as opening them to gaming from bad actors.[60]

Experimental

Three studies seek to create an experimental condition that test a treatment against a control group. Wolfowicz et al. conduct a study of 96 young males in East Jerusalem that, prior to the study, did not use Twitter.[61] They test whether there was an interactive relationship between filter bubbles, echo chambers, and the justification of violence. The treatment group suppressed algorithms by signing up to a new Twitter account with a new email and rejected all the platforms' automated recommendations. The control group used existing emails and accepted the recommendations. They tested for an echo chamber effect using a range of social network variables. As well as using data from Twitter, they also asked the individuals survey questions, such as whether they felt that suicide bombings were ever justified. This can be considered an example of an externally accessed study that can manipulate the recommendation algorithm to test the potentially harmful effects on users, although it still does not have access to the inner workings of the algorithm.

Whittaker et al. create experimental conditions on YouTube and Reddit by creating three accounts and following the same set of channels or subreddits (10 far-right; 10 neutral).[62] The accounts were then left dormant for a week so the recommendations could be collected twice per day without any interaction to create a baseline. Then, each of the accounts was subjected to one treatment: one interacted primarily with far-right channels or content, one interacted primarily with neutral content and one continued to do nothing. Each of these treatments also lasted for one week. This offered a basis of comparison both against the baseline (i.e. what has changed from the beginning?) and the control groups (i.e. how has interaction with far-right content changed compared to other alternatives?).

Finally, Huszár et al. create a randomized natural experiment on Twitter.[63] When Twitter introduced a machine-learning personalized timeline in 2016, it excluded 1% of its users, who instead see content in chronological order. This latter group acted as a control, allowing the researchers to compare whether certain types of politicians (including far-right and far-left) or media source had greater algorithmic amplification (i.e. if their tweets reached a higher proportion of personalized timelines than chronological ones).

Tracking User Behaviors

Two studies track how users act online. Attempting to understand echo chambers and recommendations on YouTube, Hosseinmardi et al. use longitudinal data from a US nationally-representative sample of over 300,000 users from Nielsen's desktop web panel from 2016 until the end of 2019.[64] Their research objective is to investigate whether YouTube's recommendation system systematically directs users to far-right content.

59 Michael A. DeVito, "From Editors to Algorithms," Digital Journalism, 5, no. 6 (2017): 753–773.

60 Nicholas Diakopoulos, "Algorithmic Accountability: Journalistic investigation of computational power structures," Digital Journalism, 3, no. 3 (2015): 398–415.

61 Wolfowicz, Weisburd, and Hasisi, "Examining the interactive effects."

62 Whittaker et al., "Recommender Systems."

63 Huszár et al., "Algorithmic Amplification."

64 Hosseinmardi et al., "Evaluating."

To do this, they examined how much of overall consumption was made up of news-related content; whether far-right channels had increased over the research period; the pathways toward far-right channels (i.e. from recommendations or not); and whether longer session times led to more extreme content.

Chen and colleagues use a similar approach in coordination with a nationally-representative survey.[65] Firstly, participants conducted the Cooperative Congressional Election Survey, which asked a range of questions about politics, society, and values. Then, a sample of the participants was asked to install an extension that tracks their activity on YouTube (including which recommendations were shown and engaged with). The combination of survey and tracking data allowed the researchers to explore whether individuals with existing beliefs – such as racial resentment – were more likely to be shown extreme content than those who do not hold such beliefs.

API-Mining

The most common approach employed by the studies under review is accessing a platform's Application Programming Interface (API) to view content that could be shown as part of recommendations. O'Callaghan et al. utilize a set of seed channels on YouTube to acquire metadata for 1000 videos, which were randomly sampled for up to 50 videos per seed.[66] They then acquired metadata for the top ten Related Videos (i.e. videos that would be recommended). The metadata for the channels of these videos were then retrieved in the same way – metadata for 1000 videos and then up to 50 were randomly sampled. These data were then analyzed to establish the extent to which the related channels for a far-right seed feature extremist content.

Ribeiro et al. also use the YouTube API to conduct a large-scale audit of what they call "user radicalization" by identifying 360 channels which are categorized into three groups – "Alt-Right," "Alt-Light," and "Intellectual Dark Web" – as well as a control group of popular media channels.[67] Similar to the O'Callaghan study, they identify the related channels – 2.47 million videos from 14,283 channels – and run "random walker" simulations to identify the navigation between these channels. Schmitt et al. take this approach too, utilizing two counter-messaging campaigns on YouTube and then using the API to collect data on the related videos with the aim of assessing the closeness of the content to extreme material online.[68] They then drew a randomized sample of 30% of all of the videos which were analyzed qualitatively, and a network analysis was conducted on the two datasets.

Ledwich and Zaitsev utilize both the YouTube API and a scraper to collect data on 816 channels spanning both mainstream and extreme content which are grouped into an ideological category. The study retrieves the impressions that the recommendation algorithm provides users of the channels.[69] They then test four hypotheses:

1. That recommendations influence viewers of radical content to view further radical content;
2. That the algorithm favors [mainstream] right-wing content;
3. The recommendation algorithm exposes users to more extreme content than they would otherwise seek

65 Chen et al., "Exposure to Alternative & Extremist Content."

66 O'Callaghan et al., "Down the (White) Rabbit Hole."

67 Ribeiro et al., "Auditing Radicalization Pathways."

68 Schmitt et al., "Counter-messages as prevention."

69 Ledwich and Zaitsev, "Algorithmic Extremism."

out; and

4. The algorithm promotes a pathway from center-right or center-left to their respective extremes.

Papadamou et al. derive a set of 6,500 "incel-derived" videos that are outlinked from a range of subreddits as well as drawing a sample of 5,700 videos as a baseline for comparison. The data are accessed via YouTube's API and coded using a lexicon of incel-related words to examine videos' transcripts, metadata, and comments. To test whether the recommendation system contributes towards steering users toward incel communities, the researchers run "random walker" simulations to explore how a user could move from one video to others.[70]

Murthy identifies a seed of 11 ISIS videos that were active on YouTube in 2016.[71] He then queries the API to establish a network of the potential (a) recommended videos for the seed, (b) those recommended by the recommended videos, and (c) videos recommended by recommended videos in (b). From this he establishes a network of over 15,021 nodes and 190,087 "recommended edges." He also collects various metadata, including genre labels, view count, and comment counts, with the aim of determining whether ISIS videos were being recommended, and if so, which types of videos were recommending them. He supplements this quantitative analysis with a qualitative comparative analysis to establish whether he could identify a set of attributes that might help explain YouTube's recommendation algorithm's decision-making process.

An important point to note is that the studies discussed above which seek to find YouTube's Related Videos as a proxy for the recommendation system cannot take user personalization into account. The studies by Ribeiro et al., Ledwich and Zaitsev, and Papadamou et al. all highlight this as a limitation.[72] Instead, they provide a snapshot of what the system could recommend based on how YouTube categorizes channels. Papadamou and colleagues attempt to simulate personalization by running a random walk after watching a few incel-related videos to mimic what a user beginning to become involved in the community might do.[73] Murthy notes that he made a concerted effort not to return personalized results by using The Onion Router (TOR) browser so results would not be biased based on cookies, location, or IP address.[74]

The investigation by Whittaker et al. involved accessing Gab's API and observing the platform's different timelines, "Recent," "Popular," and "Controversial."[75] They judge the latter two to be driven by recommendation algorithms, but the former to be based on time. By comparing the timelines, they assess whether content on the "Popular" or "Controversial" timelines were more extreme than the organic flow of posts. Unlike their research on YouTube and Reddit, this does not constitute an experiment, merely a comparison between different sources of content.

Gaudette and colleagues analyze Reddit's 'upvoting' algorithm by extracting the 1000 most popular posts in the

70 Papadamou et al., "How over is it?"

71 Murthy, "Evaluating Platform Accountability."

72 Ribeiro et al., "Auditing Radicalization Pathways"; Ledwich and Zaitsev, "Algorithmic Extremism"; Papadamou et al., "How over is it?."

73 Papadamou et al., "How over is it?"

74 Murthy, "Evaluating Platform Accountability."

75 Whittaker et al., "Recommender Systems."

"r/The_Donald" subreddit using Reddit's API.[76] They compared these to a random sample of 1000 posts. These posts were then categorized into broad themes such as "internal threat" and "external threat". Because upvoted content is pushed to the top of users' screens, they judge whether the algorithm promotes hateful content compared to a random sample.

4.6 Other Approaches

Other studies take a more qualitative or observational approach. Berger takes his readers through several steps, beginning with starting a new account on Twitter, following the account of al Nusra Front, and observing the site's "You Might Also Want to Follow" recommendations.[77] It would be fair to say that this piece is written more as an op-ed, yet it is included as it utilizes primary empirical evidence. As mentioned above, Waters and Postings do not seek to analyze the role of Facebook's recommendation algorithms, but in conducting their research observe the site's "Recommended Friends" algorithm.[78] Baugut and Neumann take a different approach, conducting 44 in-depth interviews with German and Austrian Islamists to explore the media diet and circumstances of their participants, which yielded findings that relate to platforms' recommendations.[79] The latter approach offers an important perspective because it puts content-sharing algorithms in the context of a wider media diet rather than focusing on them in isolation.

Coding

Account Categorization

Many of the studies deal with datasets that are too large to manually code each piece of content, instead deciding to categorize the channel from which the content comes. O'Callaghan et al. develop a set of far-right themes from the academic literature (e.g. anti-Islam; neo-Nazi) and use the retrieved text metadata to categorize channels, which were checked for reliability against Freebase, a topic annotation service provided by YouTube.[80] The Ribeiro et al. study, which collects over 300,000 videos, also categorize by channel, which are coded manually into either "Alt-Right," "Alt-Lite," or "Intellectual Dark Web."[81] This was done by collecting seed channels from the academic literature, which were independently annotated twice and disregarded if there was disagreement. For the recommendation dataset, they code the channels by having two experienced raters independently categorize the channels with 75% agreement. Where the coders disagreed, they discussed the cases until they reached consensus.

Ledwich and Zaitsev categorize channels using "soft tags" such as "Conspiracy," "Revolutionary," and "Partisan Right," as well as "hard tags" which differentiated between mainstream media sources and independent YouTubers.[82] The data is then coded by three labelers and a majority was needed to assign a categorization,

76 Gaudette, Scrivens, and Davies, "Upvoting Extremism."

77 Berger, "Zero Degrees."

78 Waters and Postings, "Spiders of the Caliphate."

79 Baugut and Neumann, "Online propaganda use."

80 O'Callaghan et al., "Down the (White) Rabbit Hole."

81 Ribeiro et al., "Auditing Radicalization Pathways."

82 Ledwich and Zaitsev, "Algorithmic Extremism."

with most of the values receiving an interclass correlation coefficiency that is deemed "fair" or better. Hosseinmardi et al. categorize their data into five political categories (far-left; left; center; right; and far-right),[83] drawing from the coding results outlined by Ledwich and Zaitsev and Ribeiro et al. above. For their study on Twitter, Huszár et al. also use a scale to categorize political partisanship; parties were determined as being "far-right" or "far-left" if their Wikipedia entries mentioned an association with far-left or far-right ideologies, or if the 2019 Chapel Hill Expert Survey indicated that the party was extreme (with a score above nine or below two).[84]

For their study on viewing behaviors on YouTube, Chen and colleagues draw from existing academic literature to identify 322 "alternative" and 290 "extremist" channels.[85] These sources included Becca Lewis' Alternative Influence report,[86] Ledwich and Zaitsev, Ribeiro et al., the Anti-Defamation League's Centre on Extremism, the Counter-Extremism Project, the Southern Poverty Law Centre, Hope Not Hate, as well channels found on the white supremacist website Stormfront.

Mixed-Method Content Classification

Some studies integrate a qualitative coding element to a portion of the content, which is then scaled up. Schmitt et al. drew a random sample of 30% of their whole dataset which was then qualitatively analyzed into categories that reflect non-extreme content (e.g. entertainment, news & politics, gaming) as well as themes such as "conspiracy theories," "hate speech," and "far-right" and "Islamist" propaganda.[87] To address the subjectivity of the coding process, the raters discuss divergent opinions until they are resolved.

Papadamou and colleagues create a lexicon of 200 incel-related words to cross-reference against YouTube videos' transcripts, metadata, and comments.[88] This is done by crawling the glossary on the incels.wiki webpage (resulting in 395 words) and then qualitatively removing general-purpose words (e.g., fuel, hole, legit, etc.). The annotators only included a word if they believed it was relevant, if it expresses hate or misogyny, or is directly associated with incel ideology. The coders worked independently and scored a Fleiss' Kappa of 0.69. They then selected a random sample of 1,000 videos that had been derived from incel subreddits and the first author manually annotated as either "incel-related" or "other." They counted the number of incel terms in the manually annotated transcript and comments, then used this as a base to automate the rest of their analysis.

Murthy begins by collecting seed ISIS videos by attribution to its media wing Al Hayat Media Center.[89] These 11 seeds led to a subset of 67 videos that recommended the seeds which are manually coded as belonging to ISIS if the group claimed the video officially and it had the group's logo. He then automates the process for the wider subset of 15,021 videos via keywords (e.g. "mujatweets," "ISIS/IS/ISIL" or an official video title). If the automated script finds a keyword, it is flagged and checked by a human researcher. Murthy then used qualitative comparative analysis to iteratively attempt to better understand the decision-making of YouTube's algorithms when recommending ISIS content.

83 Hosseinmardi et al., "Evaluating."

84 Huszár et al., "Algorithmic Amplification."

85 Chen et al., "Exposure to Alternative & Extremist Content."

86 Rebecca Lewis, "Alternative Influence: Broadcasting the Reactionary Right on YouTube," (2018) https://datasociety.net/research/media-manipulation.

87 Schmitt et al., "Counter-messages as prevention."

88 Papadamou et al., "How over is it?"

89 Murthy, "Evaluating Platform Accountability."

Qualitative Content Coding

Other studies that utilized smaller datasets coded the content qualitatively in its entirety rather than the channels from which it originates. Whittaker et al. utilize the Extremist Media Index,[90] which was developed by Holbrook,[91] which categorizes content on three levels: Moderate, Fringe, and Extreme, with the latter level subdivided into four levels depending on the specificity of the threat of violence. In their study, two coders worked on a random sample of 105 pieces of content, which yielded an agreement of 80% (for a Krippendoff's alpha of 0.74). After identifying the 1000 most "upvoted" posts in r/The_Donald and a random sample of 1000 posts, Gaudette et al. also code their content line-by-line descriptively which were eventually grouped into larger qualitative categories using a thematic analysis such as "external" and "internal" threat.[92] Baugut and Neumann, who conduct interviews, utilize an interpretive qualitative content analysis.[93] They included quality checks such as "red flagging" when the authors believe a participant may have been being untruthful or had internal contradictions, as well as utilizing categories of propaganda from their bespoke radicalization model.

No Coding

Other studies did not appear to have any formal coding. Given their observational nature, the Berger[94] and Waters and Postings[95] studies do not appear to have conducted any formal coding. In their study, Wolfowicz et al. do not code online content as extremist or not, but instead use surveys to gauge participants' support for suicide bombing, as well as Twitter API data to assess their online social networks.[96]

Coding "Extremism"

As noted in the introduction, concepts such as "extremist" and many related terms (such as radicalization, terrorism, far-right, etc.) are essentially contested concepts.[97] In essence, there are no agreed-upon definitions of these words and each of them is value laden. Macdonald and Whittaker argue that the lack of conceptual clarity is particularly problematic in extremism research for three reasons: it affects the robustness of empirical research, it impedes the ability to conduct meta-reviews (such as this one), and it is difficult to articulate findings to interested parties.[98]

This conceptual ambiguity can be seen within this corpus. Although all the studies that are included in this review mention extremist content or related concepts, the approaches to coding do not all seek to classify content as "extremist," but often instead use proxy terms. For example, Ribeiro et al. frame their research as auditing

90 Whittaker et al., "Recommender Systems."

91 Donald Holbrook, "Designing and Applying an 'Extremist Media Index,'" Perspectives on Terrorism 9, no. 5 (2015): 57–68.

92 Gaudette, Scrivens, and Davies, "Upvoting Extremism."

93 Baugut and Neumann, "Online propaganda use."

94 Berger, "Zero Degrees."

95 Waters and Postings, "Spiders of the Caliphate."

96 Wolfowicz, Weisburd, and Hasisi, "Examining the interactive effects."

97 Gallie, "Essentially Contested Concepts."

98 Stuart Macdonald and Joe Whittaker, "Online Radicalization: Contested Terms and Conceptual Clarity" in Online Terrorist Propaganda, Recruitment, and Radicalization, ed. John Vacca (Boca Raton: CRC Press, 2019): 33–46.

"radicalization pipelines," but do so by classifying pathways among the intellectual dark web, the alt-light and alt-right.[99] This further stretches the ambiguity because each of these terms is contested: "Identifying such communities and the channels which belong to them is no easy task: the membership of channels to these communities is volatile and fuzzy, and there is disagreement between how members of these communities view themselves and how they are considered by scholars and the media."[100]

Many studies seek to classify content as "extremist" (or a related concept) by categorizing the source, such as the channel or account which produces it. This has some clear limitations. As articulated by several of the studies in this approach, the "filter bubble" or "radicalization pipeline" hypothesis is that recommender systems steer users towards more extreme content.[101] By coding channels rather than content, authors are not able to identify whether the content that users are being steered towards is actually extreme or not. Rather, to code a channel as extreme and then analyze a corpus of videos carries an assumption that every piece of content that a channel produces is equally problematic. Although the lexicon-based approach adopted by Papadamou and colleagues offers a novel solution to this problem, it has a similar issue in that it assumes that all words related to incel ideology are extreme, noting that coders were to consider a term relevant if it expressed hate, misogyny, or is directly associated with incel ideology.[102] One of the examples they offer is "Beta male," which while certainly can be used in an extremist context, has a considerably wider usage within popular culture.

An interrelated issue is classifying channels based on existing literature and online databases. Several of the studies drew from other research, such as Lewis' or the Anti-Defamation League's reports.[103] However, these original reports did not necessarily attempt to define or identify extreme content; rather, each of them analyzed or described different aspects of the contemporary far-right. Similarly, some drew from other studies in this corpus: Chen and colleagues use both Ribeiro et al. and Ledwich & Zaitsev's classifications to inform their categories of "alternative" or "extreme," even though these were not labels that the original authors used.[104] On one hand, it is good to draw from existing scholarly work to inform the research design of an empirical project, but this may come with a limitation of not being fully aligned with the original authors' intentions.

These conceptual issues are compounded when considering the legality of the content under study, which in turn affects social media platforms' obligations to remove it. 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.[105] This is particularly stark given the different transatlantic approaches to free speech, with the US providing substantially more protection.[106] Perhaps because of this complex international environment, studies in this corpus tend not to focus on whether content is illegal or

99 Ribeiro et al., "Auditing Radicalization Pathways."

100 Ribeiro et al., "Auditing Radicalization Pathways," 3.

101 For example, see Ribeiro et al., "Auditing Radicalization Pathways"; Ledwich and Zaitsev, "Algorithmic Extremism."

102 Papadamou et al., "How over is it?"

103 Lewis, "Alternative Influence"; "From Alt Right to Alt Lite: Naming the Hate," Anti-Defamation League, 2019, https://web.archive.org/web/20190422202936/https://www.adl.org/resources/backgrounders/from-alt-right-to-alt-lite-naming-the-hate.

104 Chen et al., "Exposureto Alternative & Extremist Content."

105 Chris Meserole and Daniel Byman, "Terrorist Definitions and Designations Lists Key Findings and Recommendations," Global Research Network on Terrorism and Technology 7 (2019).

106 Peter Neumann, "Options and Strategies for Countering Online Radicalization in the United States," Studies in Conflict and Terrorism, 36, no. 6 (2013): 431–459.

not, although in some cases it can be inferred: Murthy's study is focused explicitly on ISIS videos,[107] while Berger's piece mentioned the promotion of al-Qaeda accounts,[108] both of which are designated as terror organizations in the vast majority of national lists.[109] However, when reviewing the studies that do offer either a full or partial list of channels, it is likely that the classification of "extreme" (or related concepts) falls under what the EU Counter-Terrorism Commissioner calls "legal but potentially harmful content… [which may] bring some people to embrace violent extremism."[110] This picture becomes even more complicated when considering that many studies classified by channel rather than content.

Findings

YouTube

As noted above, YouTube is by far the most popular platform under study in this corpus. Below, the results of the research on YouTube are divided into three categories: those that find the platform's content-sharing algorithms recommend extremist content; those with mixed or equivocal findings; and those that suggest that there is no "filter bubble" effect.

Positive effects

After categorizing channels, the O'Callaghan et al. study finds that there was support for the hypothesis that YouTube's algorithm recommended further far-right content, and in turn, could result in users being excluded from information that is not already in line with their ideological perspective.[111] They describe this as an "immersive ideological bubble" and argue that it places an emphasis on platforms as important political actors, whose recommendation systems are not neutral in their effects. Similarly, in their research on YouTube, Whittaker et al. find that the treatment which interacted primarily with far-right content is significantly more likely to recommend further content classified as both "Extreme" and "Fringe."[112] Conversely, the account which primarily interacted with neutral content is significantly less likely to do so, as was the baseline account. In essence, both studies show that YouTube's content-sharing algorithms are reactive to viewing extremist content and will recommend it further.

In their interview-based research on the media diet of Islamists, Baugut & Neumann find that many individuals began with a basic interest in Islam or in news media that was outside the mainstream. These individuals then followed platforms' algorithmically influenced recommendations to where they encountered radical propaganda. One incarcerated interviewee made this point:

I've never searched for it, but when you type 'Islam' on YouTube, you automatically get videos from Islamists. Or if you are looking for a preacher, then all the other [i.e., radical] preachers will also come, they have the

107 Murthy, "Evaluating Platform Accountability."

108 Berger, "Zero Degrees."

109 "Group Inclusion Policy," Tech Against Terrorism, (n.d.), https://www.terrorismanalytics.org/group-inclusion-policy.

110 Council of the European Union, "The Role of Algorithmic Amplification in Promoting Violent and Extremist Content and its Dissemination on Platforms and Social Media," (Brussels, 2020): 2.

111 O'Callaghan et al., "Down the (White) Rabbit Hole."

112 Whittaker et al., "Recommender Systems."

same topics and then comes ISIS.[113]

Respondents also note that if they liked nasheed music on YouTube, then they were also recommended accompanying videos that depicted violence. Another interviewee highlights that he was propelled to act because of violent propaganda offered by YouTube, which propagated a victimhood narrative. Overall, these studies highlight these algorithms as an important and problematic factor in their participants' radicalization.

Mixed effects

Schmitt et al. find that extremist content may exist in close proximity to YouTube counter-message campaigns.[114] Both the seed accounts that they begin with (#WhatIS and ExitUSA) differed in the amount and diversity of extremist content to which they relate, which they put down to structural differences between the two campaigns. ExitUSA is mostly connected with a diverse environment of entertainment and information-related videos and only a small amount of far-right propaganda. However, they note that users can still be easily confronted with this content within two clicks via recommendations. On the other hand, some of the #WhatIS videos have a remarkable number of connections with Islamist propaganda, which they explain by the thematic overlap of specific keywords (such as "jihad") which the algorithm was assigning similarity. They note that it could be argued that extreme videos connected in the same networks as counter-messaging is positive as it gives an opportunity for the latter to reach those who are engaging with radical propaganda. However, they also note that extremists tend to be much more prolific than counter-message creators. Overall, they urge caution when crafting counter messages on social media because they may have an unintended consequence of pushing individuals closer to problematic content.

Ribeiro et al. analyze over two million YouTube recommendations that were related to their dataset of three categories: Alt-Right, Alt-Lite, and Intellectual Dark Web.[115] They find that YouTube's recommendation algorithm frequently suggests Alt-Lite and Intellectual Dark Web content, and once in these communities, it is possible to find the alt-right from recommended channels – but importantly, not from recommended videos. They qualify these findings by stating they were only able to sample a small proportion of the total recommendations and by noting that they were unable to account for personalization. Despite these limitations, they argue that their findings support the notion that there is a "radicalization pipeline" on YouTube.

In Murthy's dataset, which was collected and curated in 2016, he suggests that the chances of users finding ISIS content on YouTube are rare, but not zero.[116] He finds that when such content was recommended to users, it tended to either be from other ISIS videos, or when important metadata was shared, such as title keyword similarity – supporting the findings of Schmitt et al.[117] Importantly, he finds that the language of the video played an important role in potential recommendations; under 12% of the videos that recommended the seeds were English, while two-thirds were Arabic. This suggests that, as noted above, if research is focused on English-language content, it may skew results and miss key factors.

113 Baugut and Neumann, "Online propaganda use," 1576.

114 Schmitt et al., "Counter-messages as prevention."

115 Ribeiro et al., "Auditing Radicalization Pathways."

116 Murthy, "Evaluating Platform Accountability."

117 Schmitt et al., "Counter-messages as prevention."

Papadamou and colleagues' research on incels finds that the community steadily increased over their timeframe of data collection.[118] They find that there is a small but "non-negligible amount" of incel-related videos (2.9%) within YouTube's recommendation graph recommended to users. However, their random walks suggest that if a user begins to watch incel-related videos, the algorithm recommends other incel-related content with increasing frequency. Similarly, Chen et al. also find a small but non-zero number of "extreme" or "alternative" videos recommended to its participants; over 98% of total recommendations did not lead to these types of videos.[119] However, they find that when users did watch these categories of videos, their chances of being recommended further ones increased; watching an "alternative" video led to 37% of the recommendations being further videos in this category and 2.3% being "extreme" videos. Moreover, when users watched an "extreme" video, 29% of the recommendations were to other channels in this category and 14% to an "alternative" channel. These findings are similar when considering user behavior; 98.8% of all followed recommendations were to non-extremism and non-alternative channels. However, around half of the followed recommendations from alternative or extreme videos were to further channels in their respective categories. When cross-referencing against the self-report survey, those that watched these channels were almost entirely made up of individuals that reported high levels of racial resentment, suggesting that user choices may play a bigger role than recommendation systems.

Minimal-to-no effects

The findings of one study, authored by Ledwich and Zaitsev, suggest that YouTube recommendation algorithms actively discouraged users from visiting extreme content online, which they argue refutes popular "radicalization" claims.[120] As noted above, they test four hypotheses. The first is that recommendations influence individuals to watch more content than they would otherwise have and reduce the number of alternative views. They find partial support for this: while there is a clear preference from categories towards other channels in the same category, they do not find evidence that there is a dramatic shift from extreme content to further extreme channels. Secondly, they find no support that there was a right-wing advantage within YouTube's recommendation system, instead suggesting that mainstream news is preferred. The third hypothesis is that the algorithm pushes users towards more extreme content than they would otherwise have seen. They also find no support for this; their data suggests that the algorithm actually appears to restrict traffic towards extreme right-wing categories. The final hypothesis is that there is a far-right radicalization pathway that takes users from the center ground, via content that is increasingly critical of left-wing or centrist narratives, towards the extreme right. Their findings suggest that mainstream right-wing news channels such as Fox News benefit from the recommendation algorithm, but that smaller fringe YouTubers were disadvantaged, causing the authors to reject the hypothesis. Overall, Ledwich and Zaitsev reject the notion that the platform's recommendation system is a radicalization pipeline.

It is worth noting that this study has been the subject of some critique, with Ribeiro and colleagues arguing that the authors make three key mistakes.[121] Firstly, the experiment does not take personalization into account, which means that the recommendation system is more likely to suggest mainstream channels, which would not

118 Papadamou et al., "How over is it?"

119 Chen et al., "Exposureto Alternative & Extremist Content."

120 Ledwich and Zaitsev, "Algorithmic Extremism."

121 Manoel Horta Ribeiro et al., "Comments on 'Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization,'" UCL Information Security, December 29, 2019, https://sec.cs.ucl.ac.uk/posts/2019/12/youtube-radicalization-study/.

necessarily be repeated for an actual user. Secondly, the authors collected data after a major policy change (as discussed above), but they treat the findings as if they were from a previous version. Finally, the paper lacked a robust methodological explanation, such as including limitations (and adjusting them to their findings), reporting statistical significance, or providing clarity for their inter-coder reliability tests. Zaitsev responded to this critique, arguing that using an anonymous account is not a major flaw (and it was discussed in their limitations), clarifying that their claims only relate to YouTube post-policy changes, and defending their methodological transparency.[122]

The study of user preferences on YouTube by Hosseinmardi et al. also downplays the role of recommendation systems. They find that the total amount of far-right content increased over the four years under study, and those that consume it tend to show a more extreme pattern of engagement compared to those with other ideologies.[123] Importantly, they find that users self-segregate into ideological communities when watching content – i.e. an "echo chamber" effect. However, they find that the pathways that channel users towards far-right videos are diverse and much of the pathway comes from other platforms, arguing that only a fraction can be attributed to recommendations. They also find that longer sessions do not lead to more extreme content. They note that YouTube may be a source of concern given that extreme content is discoverable. However, they argue that the focus on recommendation systems is too narrow, and it is better to consider the platform as one part of a wider ecosystem. They also argue that users should be viewed as active participants, purposefully seeking out content rather than being passively recommended.

Reddit and Gab

Gaudette and colleagues' study explores whether Reddit's upvoting and downvoting algorithm facilitates "othering" discourse on the subreddit r/The_Donald. They compare the 1000 most upvoted posts – which are more likely to be shown to users – to a random sample of 1000 posts. They find that the upvoted sample contained substantially more extreme discourse than the random one, which they group into two themes.[124] First was mention of the "external threat," which comprises 11.6% of the upvoted sample, made up of hateful comments about Muslims. By comparison, only 1.6% of the random sample contains this type of content. Secondly, the "internal threat" is prevalent in 13.4% of posts in the upvoted sample, which targets the left as a violent enemy that wants to hurt the West. This is compared to 5.7% in the random one. The authors note that for both groups, the random sample not only contains fewer references to the respective out-group, but the language is less extreme as well. They argue that their findings suggest that Reddit's upvoting and downvoting algorithm plays a key role in "othering" the out-groups and in turn facilitates a collective extreme identity on r/The_Donald.

Although the YouTube portion of the study by Whittaker et al. finds the recommendation system to promote extreme content, their investigations on Reddit and Gab found no such relationship.[125] Despite there being 30 pieces of content (1.4%) that were classified as "extreme" on Reddit and 416 (20%) that were judged to be

122 Anna Zaitsev, "Response to Further Critique on our Paper 'Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization,'" Medium, January 8, 2020, https://anna-zaitsev.medium.com/response-to-further-critique-on-our-paper-algorithmic-extremism-examining-youtubes-rabbit-hole-af3226896203.

123 Hosseinmardi et al., "Evaluating."

124 Gaudette, Scrivens, and Davies, "Upvoting Extremism."

125 Whittaker et al., "Recommender Systems."

"fringe," interacting with either type of content does not make the platform more likely to promote it more in future. They do find that interacting with neutral content decreased the likelihood of being recommended "fringe" content, suggesting that there may be some filtering effects. With regards to Gab, there were no significant differences between the "Latest," "Controversial," or "Popular" timelines and the amount of "Extreme" content. They do find that "Fringe" content may be prioritized in the "Popular" timeline. Overall, the authors argue that there is no evidence that either of these platforms promote extreme content via the algorithm.

Twitter

In their experimental study on new Twitter users, Wolfowicz et al. find evidence to support their hypothesis that there is an interactive relationship between filter bubbles, echo chambers, and justification of suicide bombing.[126] In each of their models, the treatment (suppressing algorithmic control on Twitter) is not found to directly affect radicalization. However, when considering the interactions between their proxy variables for echo chamber effects, they find an interactive relationship. For example, they find that for individuals in the treatment group that had more outwardly focused networks, there was a decreased likelihood of them justifying terrorism. In essence, the findings suggest that the filter bubble alone does not cause radicalization, but it could be a contributing factor, particularly when considered alongside particular types of networks.
In their randomized natural experiment on Twitter, Huszár et al. find no evidence that far-right or far-left accounts are amplified on users' timelines.[127] They find that mainstream right-wing political parties and right-wing news sources are amplified because users with an algorithmically driven timeline are significantly more likely to see these sources than those with a chronological one. However, they also find that in countries where there are a substantial number of elected officials that can be described as far-left or far-right (such as the Japanese Communist Party; the AfD in Germany; or VOX in Spain), the amplification is lower for these parties than for centrist parties in the same countries.

Account Recommendations

Two studies suggest that platforms may recommend extreme accounts as suggested user connections. Berger's observation of Twitter's "Who to Follow" suggestions shows that if a new user follows the now-suspended account for Syrian group al Nusra Front[128] – which was at the time associated with al-Qaeda but has since merged into Tahrir al-Sham[129] – the user was then recommended the account for the radical Ansar al-Mujahideen forum. If the user follows this account, then Twitter suggests further prominent jihadists. Importantly, the "neutral" default accounts that Twitter typically offers new users such as Justin Bieber and Lady Gaga become less frequent and are replaced by "hardcore terrorists and extremists." He also notes that the same effect can be seen if the user follows far-right accounts instead like the American Nazi Party. Waters and Postings found that "Facebook's algorithms have also actively helped connect IS supporters and build extremist networks through 'suggested friends.'"[130] They found that it was the most likely explanation for connecting two of the supporters in their sample and both of the authors discovered their own accounts recommended by

126 Wolfowicz, Weisburd, and Hasisi, "Examining the interactive effects."

127 Huszár et al., "Algorithmic Amplification."

128 Berger, "Zero Degrees."

129 Europol, "TE SAT: European Union Terrorism Situation and Trend Report," (The Hague, 2018).

130 Waters and Postings, "Spiders of the Caliphate," 78.

jihadists after engaging with other extreme individuals. Both Berger and Waters and Postings argue that the respective platforms are inadvertently creating a network that helps to connect extremists because they desire connectivity of individuals with similar interests.

Discussion

Findings and Knowledge Gaps

Looking at the corpus of studies as a whole, the findings suggest that content-sharing algorithms may amplify extreme content towards users. Of the fifteen identified studies, only two conclude that platforms had minimal effects,[131] while one found that they actively push users away from extreme content.[132] One study had split results on different platforms – i.e. YouTube does amplify extreme content, but Reddit and Gab do not,[133] while one study found no direct effect but an interactive relationship with other variables.[134] The other ten suggested that, in various ways, recommendation systems can promote extreme content, although as noted above, this often comes with important qualifications.

Understanding the types of data that were utilized in this corpus offer an insight into the gaps in the existing literature. YouTube is by far the most researched platform within the sample, which from one perspective is intuitive given the high level of media coverage of the platform – for example, the Christchurch killer was described by news outlets as being radicalized on YouTube.[135] However, the platform also has a researcher-friendly API which makes it more feasible to access datasets and explore the role of recommendations than other platforms.[136] In comparison, we have little empirical data as to whether Facebook algorithms promote hateful or extreme content, which is concerning given the recent testimony of whistleblower Frances Haugen, who suggests they do based on leaked internal research.[137]

More broadly, the corpus contains just five platforms, and only one can be described as "small" (Gab). Terrorists and extremists populate a range of large and small platforms, often simultaneously, to adapt to the hostile environment in which they find themselves.[138] Tech Against Terrorism find that ISIS used up to 330 different platforms,[139] while other studies (on both jihadist and far-right content) have found outlinks from Twitter to

131 Hosseinmardi et al., "Evaluating"; Huszár et al., "Algorithmic Amplification."

132 Ledwich and Zaitsev, "Algorithmic Extremism."

133 Whittaker et al., "Recommender Systems."

134 Wolfowicz, Weisburd, and Hasisi, "Examining the interactive effects."

135 Sam Shead, "YouTube radicalized the Christchurch shooter, New Zealand report concludes," CNBC, December 8, 2020, https://www.cnbc.com/2020/12/08/youtube-radicalized-christchurch-shooter-new-zealand-report-finds.html.

136 Whittaker et al., "Recommender Systems."

137 C-SPAN, "Facebook Whistleblower Frances Haugen testifies before Senate Commerce Committee," YouTube, October 5, 2021, https://www.youtube.com/watch?v=GOnpVQnv5Cwandab_channel=C-SPAN.

138 M. Meili Criezis, "Remaining and Expanding or Surviving and Adapting? Extremist Platform Migration and Adaptation Strategies," Global Network on Extremism and Technology, 2021, https://gnet-research.org/2021/11/12/remaining-and-expanding-or-surviving-and-adapting-extremist-platform-migration-and-adaptation-strategies/.

139 "ISIS use of smaller platforms and the DWeb to share terrorist content summary," Tech Against Terrorism, April 29, 2019, https://www.techagainstterrorism.org/2019/04/29/analysis-isis-use-of-smaller-platforms-and-the-dweb-to-share-terrorist-content-april-2019/.

dozens of different platforms.[140] While it remains vital to understand how extremists exploit the largest platforms (as research demonstrates a complex online ecosystem in which the largest platforms still play an important role[141]), an emphasis on larger ones may lead to a knowledge gap.

The content being mostly English-language or Western-focused also may be an important gap in our understanding of potential amplification. This is likely related to the proximity or language skills of the researchers. Murthy finds that language may play an important role; Arabic videos are more likely to be recommended than English ones.[142] This may suggest that, even in an optimistic scenario in which platforms have successfully navigated the balance between removal for violative content and removing or downranking borderline content from recommendations, it could be restricted to English-language content. Platforms often trial such measures on Western audiences first; YouTube's policy of reducing borderline content was tested first in the US before being expanded outwards.[143] Murthy's findings suggest that language is an important factor and that we should not assume that policy responses have equal effects if platforms dedicate differing levels of resources to them. This is particularly important when considering the claim that Facebook's "algorithms amplified hate speech and… failed to take down inflammatory posts in the ongoing genocide against the Rohingya in Burma."[144] Importantly, the plaintiffs argue that the platform's lack of investment in local knowledge and understanding of linguistic context plays an important role. While this remains a claim, the lack of academic study in the area leaves us unable to determine the role of content-sharing algorithms.

We also lack a transparent understanding of how recommendation systems work. The studies in this corpus focus heavily on researchers externally accessing a recommendation system and then assessing its output – i.e. a "black box" approach. GIFCT's CAPPI working group note that "the limitation of such studies is that without any insight into how algorithms make recommendations, it is difficult to fully assess and understand how they may lead to different kinds of outcomes."[145] These types of approaches do not typically manipulate platforms' recommendation systems to observe their effect on users, but instead tend to manipulate real or simulated users, or merely assess the type of content that could potentially be recommended. To bridge this knowledge gap in the future, Knott and colleagues recommend empirical collaboration between external stakeholders and platforms in which internal methods can be used to study the effect of recommendation systems on users.[146] The study on Twitter by Huszár et al. is a good example of this type of collaboration,[147] but it could be expanded to attempt to understand how engaging with platforms' content – or third-party interventions – can affect

140 Maura Conway et al., "Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts," Studies in Conflict and Terrorism 42, no. 1–2 (2018): 141–160; J. M. Berger, "The Alt-Right Twitter Census," Vox Pol (Dublin, 2018).

141 Ali Fisher, Nico Prucha, and Emily Winterbotham, "Mapping the Jihadist Information Ecosystem: Towards the Next Generation of Disruption Capability," Global Research Network on Terrorism and Technology 6 (2019).

142 Murthy, "Evaluating Platform Accountability.

143 YouTube, "Our Ongoing Work."

144 Dan Milmo, "Rohingya sue Facebook for £150bn over Myanmar genocide," The Guardian, December 6, 2021, https://www.theguardian.com/technology/2021/dec/06/rohingya-sue-facebook-myanmar-genocide-us-uk-legal-action-social-media-violence.

145 "Content-Sharing Algorithms," GIFCT, 15.

146 Alistair Knott et al., "Responsible AI for Social Media."

147 Huszár et al., "Algorithmic Amplification."

behavior.[148]

An interrelated issue is that we know very little about users' choices when it comes to recommendation systems and extremist content. By their nature, most social media algorithms offer some level of user-choice (i.e., to engage with the "recommended for you" content on YouTube, or to have the personalized timeline on Twitter). One of the key debates in the broader literature on "filter bubbles" is whether they play a greater role in content selection than users' own choices.[149] This is still a largely unanswered question within the studies discussed here. Rather, most of these studies look at the "supply" of extreme content – that is to say, the environment with which users could potentially engage.[150] This leaves a causal knowledge gap as to how this may affect its audience's behaviors.

The studies which analyzed user behavior offer a mixed picture. Hosseinmardi et al. and Chen et al. both offer tentative findings which suggest that users' own choices play a bigger role than recommendation systems;[151] Wolfowicz et al. suggest an interactive relationship with users' networks;[152] and Baugut and Neumann's participants suggest that YouTube's recommendation system led them towards more extreme content.[153] To better understand whether such systems actually cause harm – rather than direct users towards potentially harmful content – research must go beyond simply investigating the environment and place more emphasis on how individuals have interacted with algorithms and the choices that they have made, as well as better understanding the feedback that they receive from the platform as a result of these choices.

Contextualizing into Wider Debates

Content-sharing algorithms do not exist in a vacuum. Instead, they are just one part of the wider radical online milieu, and we should not overlook the extremist environment on the platforms the content inhabits. Put simply, for extremist materials to be recommended (and for researchers to study this using open-source methods), it must be available on social media platforms. This point is made explicit in the study by Gaudette et al., who note that much of the content does contravene the rules of r/The_Donald, but moderators did not enforce the rules, and therefore such content was allowed to remain featured in the community.[154] On the other side of the coin, Whittaker et al. argue that their null findings on Gab demonstrate the importance of the radical community of which it is part.[155] Research has shown that the platform has become a haven for problematic far-right

148 For example, see Erin Saltman, Farshad Kooti, and Karly Vockary, "New Models for Deploying Counterspeech: Measuring Behavioral Change and Sentiment Analysis," Studies in Conflict and Terrorism, March 30, 2021.

149 For example, see Bakshy, Messing, and Adamic, "Exposure to Ideologically Diverse News,"; Ivan Dylko et al., "The Dark Side of Technology: An experimental investigation of the influence of customizability technology on online political selective exposure," Computers in Human Behavior 73 (2017): 181–190; Cédric Courtois, Laura Slechten, and Lennert Coenen, "Challenging Google Search Filter Bubbles in Social and Political Information: Disconforming evidence from a digital methods case study," Telematics and Informatics 35, no. 7 (2018): 2006–2015.

150 Ines von Behr et al., "Radicalisation in the Digital Era: The use of the internet in 15 cases of terrorism and extremism" RAND Europe, 2013.

151 Hosseinmardi et al., "Evaluating"; Chen et al., "Exposure to Alternative & Extremist Content."

152 Wolfowicz, Weisburd, and Hasisi, "Examining the interactive effects."

153 Baugut and Neumann, "Online Propaganda use."

154 Gaudette, Scrivens, and Davies, "Upvoting Extremism."

155 Whittaker et al., "Recommender Systems."

content.[156] The authors note that there was a wide range of extreme and fringe content in their dataset, which, while not algorithmically amplified, was still easily accessible to users. Discussing YouTube, Lewis argues that although recommendation algorithms could play some role in driving extreme content, the picture is substantially more complicated;[157] she argues that content is also driven by cross-networking influencers, who can provide new and large audiences,[158] and stresses the importance of the micro-celebrity status of many influencers, which can provide an illusion of authenticity to their audiences.[159]

Despite media claims of "radicalization by algorithm," we still have a very limited understanding of what part they play in the process. The distinction between radical content and users' own choices is exemplified by Munger and Phillips, who critique the idea that YouTube's recommendation algorithm plays a prominent role.[160] They argue that this understanding is little more than an update of the now-discredited "hypodermic needle model" of mass communication. Instead, they point to the affordances that YouTube offers – notably that the primary content is video and that it is a media company – has created a capacity to create radical alternative political cannons and communities to interpret them. These affordances create a "supply and demand" framework which highlights how YouTube has made content creation appealing for fringe political content creators, but also recognizes that regardless of these, they require an active audience to want to watch it. They note that the consumption of "white nationalist video media was not caused by the supply of this media 'radicalizing' an otherwise moderate audience. Rather the audience already existed, but they were constrained by the scope of the ideology of extant media."[161] In other words, like Lewis, they warn against over-interpreting research into the role of recommendation algorithms if it comes at the expense of understanding the audience.

There has been an increased policy concern about the role of recommendation algorithms in the radicalization process.[162] However, it is important to note that thirteen of the fifteen studies focused their research exclusively on the online domain. While it is intuitive to study an online effect in its own environment, it runs the risk of inflating the role of the Internet in cases of terrorism or extremism. Research has consistently shown that despite terrorists utilizing the Internet heavily, this does not come at the expense of offline interactions[163] and that offline

156 Berger, "The Alt-Right Twitter Census"; Maura Conway, Ryan Scrivens, and Logan Macnair, "Right-Wing Extremists' Persistent Online Presence: History and Contemporary Trends," ICCT Policy Brief (October, 2019); Lella Nouri, Nuria Lorenzo-Dus and Amy-Louise Watkin, "Following the Whack-a-Mole: Britain First's Visual Strategy from Facebook to Gab," Global Research Network on Terrorism and Technology: Paper No. 4, 2019.

157 Rebecca Lewis, "All of YouTube, Not Just the Algorithm, is a Far-right Propaganda Machine," Medium, January 8, 2020, https://ffwd.medium.com/all-of-youtube-not-just-the-algorithm-is-a-far-right-propaganda-machine-29b07b12430.

158 Lewis, "Alternative Influence."

159 Rebecca Lewis, "'This Is What the News Won't Show You': YouTube Creators and the Reactionary Politics of Micro-celebrity," Television and New Media 21, no. 2 (February 2020): 201–217.

160 Kevin Munger and Joseph Phillips, "Right-Wing YouTube: A Supply and Demand Perspective," International Journal of Press/Politics 27, no. 1 (January 2022): 186–219.

161 Kevin Munger and Joseph Phillips, "A Supply and Demand Framework for YouTube Politics: Introduction to Political Media on YouTube," Penn State Political Science (October, 2019): 12.

162 HM Government, "Online Harms White Paper," (London: The Stationary Office, 2019); Council of the European Union, "The Role of Algorithmic Amplification."

163 von Behr et al., "Radicalisation in the Digital Era"; Paul Gill et al., "Terrorist Use of the Internet by the Numbers: Quantifying Behaviors, Patterns, and Processes," Criminology and Public Policy 16, no. 1) (2017): 99–117; Joe Whittaker, "The Online Behaviors of Islamic State Terrorists in the United States," Criminology and Public Policy 20, no. 1 (2021): 177–203; Chamin Herath and Joe Whittaker, "Online Radicalisation: Moving Beyond a Simple Dichotomy," Terrorism and Political Violence (November 22, 2021).

social networks may play a greater role than the web.[164] Baugut and Neumann consider both domains in relation to each other. As noted above, they do suggest an important role for algorithms recommending propaganda, but also find that the online and offline domains are inseparably intertwined: "Contact with online propaganda was usually followed by personal talks with peers or preachers, and preachers used their personal talks with recruits as opportunities to show them media propaganda."[165] An over-emphasis on the online environment – including algorithms – may foster a "streetlight effect" in which research focuses on what is easily available and therefore miss the true picture of contemporary radicalization.

A final important point to consider is the potential regulation of recommendation systems. Policymakers have repeatedly signaled an intention to have greater control over how such algorithms operate. Whittaker and colleagues note that presently there is little regulation in law, and potential policy maneuvers tend to be focused on algorithmic transparency (for example, the EU's Digital Services Act or the US's Filter Bubble Transparency Act). They argue that this leaves the amplification of borderline content largely unresolved.[166] The EU Counter-Terrorism Coordinator argues that platforms should remove such content from recommendations and directly links it as being a potential conduit of radicalization.[167] However, as has been demonstrated above, this is not a link that is supported by evidence. Moreover, identifying such content is particularly problematic and undoubtedly leaves sizable grey areas. Tech Against Terrorism argue against regulatory policies of removing legal yet harmful content from recommendations, suggesting that it has negative implications for freedom of speech, the rule of law, and raises serious concerns over extra-legal norm-setting.[168]

Recommendations

1. **Broader scope:** This review demonstrates that most of the body of knowledge is drawn from research that focuses on English-language content, in Western countries, with a focus on the far-right. However, there may be good reason to believe that platforms are more finely attuned to this context at the expense of other locations, languages, and ideologies.
2. **More internal research:** To provide the clearest picture of the phenomenon, stakeholders should foster collaboration between those with access to recommendation algorithms (i.e. social media platforms) and researchers. Policymakers can aid this process by making data-sharing exemptions explicit within data protection regulations.
3. **Audit platforms' responses:** This review highlighted how platforms changed their policies on content recommendations over the last decade. However, academic research has yet to establish whether these responses are effective. For example, are authoritative voices raised into the recommendations of borderline content, or can groups tied to offline violence still be found in recommendations?
4. **Account for personalization and track user behavior:** Existing research mostly focuses on the "supply" of content that could potentially be recommended to users. Future research should utilize experimental

164 Sean C. Reynolds and Mohammed M. Hafez, "Social Network Analysis of German Foreign Fighters in Syria and Iraq," Terrorism and Political Violence 31, no. 4 (April, 2019): 661–686.

165 Baugut and Neumann, "Online Propaganda use," 1585.

166 Whittaker et al., "Recommender Systems."

167 Council of the European Union, "The Role of Algorithmic Amplification."

168 "Content Personalization and the Online Dissemination of Terrorist and Violent Extremist Content," Tech Against Terrorism, 2021, https://www.techagainstterrorism.org/wp-content/upload s/2021/02/TAT-Position-Paper-content-personalisation-and-online-dissemination-of-terrorist-content1.pdf.

designs which account for platforms' personalization rather than gathering potential recommendations. Similarly, future research should focus on "demand" by using methodological designs which track user behavior (i.e. which can show not only whether content is recommended but whether users actually follow these recommendations).

5. **Code content rather than accounts:** Many of the studies in this corpus classified accounts/channels as "extremist" (or a related term). This carries an implicit assumption that all content from the account is equally problematic. However, this is clearly not the case. If future studies' research questions relate to whether extreme content is being amplified, then it should be content that is coded.
6. **Researcher Transparency:** Researchers should offer clear explanations of the methodological decisions that are taken. This includes giving working definitions for coding, particularly when referring to essentially contested concepts, conducting and detailing inter-rater reliability tests, and providing information on how data are coded (including making datasets available if possible).
7. **Platform Transparency:** Platforms should offer an explainable rationale for why users have been recommended content. This should be available for researchers to explore empirically in future studies.

Appendix

Study	Platform	Ideology	Language	Methods	Findings
Berger (2013)	Twitter	Jihadist	Arabic	Exploration of Twitter's recommendation system. Creates new account and follows jihadist accounts.	"Who to follow" recommends a number of prominent jihadist accounts.
O'Callaghan et al. (2015)	YouTube	Far-right	English; German	Access API to draw Related Videos. Use text metadata to categorize channels, which were checked against Freebase.	Recommends further far-right content that could result in "immersive ideological bubble."
Schmitt et al. (2018)	YouTube	Jihadist; Far-right	English; German	Access API to collect Related Videos for two counter-messaging campaigns. Qualitatively analyzed and categorized 30% of dataset.	Extremist content within related videos. High crossover with anti-jihadist campaign (possible due to keyword similarity).
Waters & Postings (2018)	Facebook	Jihadist	Multiple	Social network analysis.	At least two ISIS supporters likely recommended as friends. Authors were also recommended IS-supporting accounts.
Ledwich & Zaitsev (2019)	YouTube	Far-right	English**	Access API and use scraper to collect data on seed channels. Code into categories based on ideology and mainstream vs independent.	YouTube actively discourages users from extreme content. No evidence to suggest movement towards more extreme categories.
Ribeiro et al. (2019)	YouTube	Far-right	English**	Audit seed channels which have been categorized into ideological groups. Access API to identify Related Videos and simulate navigation between channels.	YouTube recommends "Alt-Lite" and "Intellectual Dark Web" content, and once in these communities it is possible to find "Alt-Right" content, but not from recommendations. Suggest that findings support the notion of a "radicalization pipeline".
Gaudette et al. (2020)	Reddit	Far-right	English	Compare 1000 most "upvoted" posts in "r/The_Donald" against random sample.	Most upvoted sample substantially more extreme than random sample.
Baugut & Neumann (2020)	n/a*	Jihadist	German	44 interviews to explore media diet.	Individuals said that platform recommendations took them from basic knowledge to radical propaganda.

Hosseinmardi et al. (2020)	YouTube	Far-right	English	Representative sample of web users' browser history over 4 years. Channels coded according to political ideology.	Pathways towards far-right content is diverse and only a fraction can be attributed to recommendations. No trend towards more extreme content over longer sessions. Suggest user preference plays a bigger role.
Wolfowicz et al. (2021)	Twitter	Jihadist	Arabic	Recruited 96 non-Twitter users. Treatment group suppresses algorithm, control group accepts all automated suggestions. Ask participants how they feel about suicide bombing.	Interaction effect between recommendations and network effects (i.e. filter bubble and echo chamber are complementary).
Whittaker et al. (2021)	YouTube; Reddit; Gab	Far-right and male supremacist	English	YouTube/Reddit: Create identical accounts, use bot to log in engage with content. Access recommendations via API. Gab: Access data via API to compare "Recent," "Popular," and "Controversial" timelines. All: Code data according to Extremist Media Index (Holbrook 2015).	YouTube: Extreme and Fringe content more likely to be recommended and to be ranked higher. Reddit/Gab: Extreme material not promoted via recommendations.
Papadamou et al. (2021)	YouTube	Incel	English**	Compare 6.5k incel videos against 5.7k random videos. Build lexicon of 200 incel-related words to code videos as incel-related from transcript. Access via YouTube API Conduct Random Walker simulation.	Small chance that users will be recommended incel videos by system. If user watches incel-related videos, algorithm recommends other incel-related videos with increasing frequency.
Chen et al. (2021)	YouTube	Far-right	English	Link US nationally-representative survey to YouTube viewing behaviors. Identify "alternative" and "extreme" accounts via literature.	YouTube recommendations can expose users to potentially harmful content. However, vast majority of exposure is to individuals with self-reported racial resentment.
Murthy (2021)	YouTube	Jihadist	Arabic; English; French; Mandarin	Detect 11 ISIS videos as seeds. Access API to establish network that represents 1) recommended videos, 2) recommendations of recommendations, and 3) the recommendations of (2). Use qualitative comparative analysis to assess which features likely influence algorithmic decision-making.	Chance of finding ISIS content accidentally is rare, but non-zero. Usually happened when there was similar metadata. When ISIS content was recommended, it tended to be from other ISIS videos (particularly those that were not English language). Radical keywords seem to be important in recommending ISIS.

Huszár et al. (2022)	Twitter	Far-right and Far-left	English; Japanese; French; Spanish; German; Turkish	Identified 3634 legislators from seven countries. 1% of Twitter users were excluded from timeline personalization (Control group). Compare the reach of legislators from control group to treatment group who do have personalized timeline. Sought to assess whether XR/XL legislators have greater reach in personalized timelines.	No evidence to support that far-right or far-left groups are amplified more than moderate ones.

**Data derived from interviews*

*** Not explicitly stated but examples or keywords are entirely or primarily English-language*

GIFCT WORKING GROUPS OUTPUT 2022

Transparency Reporting: Good Practices and Lessons from Global Assessment Frameworks

GIFCT Transparency Working Group

GIFCT
Global Internet Forum to Counter Terrorism

Dr. Courtney Radsch
Center for Media, Data and Society

Overview

The GIFCT seeks to inform its work on transparency by examining practices and approaches from an array of sectors and stakeholders to identify good practices, lessons learned, and approaches from beyond the technology sector. There is no singular definition of transparency reporting. Rather, the concept, principles, and need for transparency are embedded in assessment frameworks covering such areas as Corporate Social Responsibility (CSR), Environmental, Social, and Governance (ESG), Environmental Impact Assessments (EIA), Social Impact Assessment (SIA), Human rights Impact Assessments (HRIA), and corporate reporting in finance/accounting/taxation, extractive industries, and information and communication technologies (ICT). The following research scoping agenda identifies key themes, good practices, and lessons drawn from these frameworks along with academic, governmental, and civil society reporting and assessments of transparency and impact reporting more broadly.[1] Following this analysis, it turns briefly to transparency reporting in the tech sector and identifies common practices and recommendations.

Reporting takes place in response to mandatory requirements as well as occurring voluntarily. A wide range of sectors and stakeholders are increasingly conducting voluntary and/or mandatory reporting on various dimensions of their work, from governance and process to activities and impacts. They are also creating sets of principles and expectations through third-party assessments, rankings, and impact reporting. Such reporting is often aimed at increasing transparency, building trust, and enabling accountability; it may also be required by law or implemented in response to external pressure. Most corporate transparency reporting is voluntary, even in highly regulated industries like the financial sector, resulting in a wide range of quality and comprehensiveness.[2]

Transparency reporting by information and communications technologies is relatively new compared to other sectors. Nonetheless, in just over a decade, regular transparency reporting by application layer tech firms has become an emergent norm,[3] though the quality, comprehensiveness, and expectations of what should be included continue to evolve even as some standards coalesce while others are contested.[4] While telecom firms and businesses that have existed in more mature sectors may have engaged in transparency or impact reporting over a longer period of time, reporting on data requests, content moderation, terms of service issues, and other issues specific to the expressive and privacy dimensions of the tech sector are still nascent.

As a relatively new organization, the GIFCT must consider how it conceptualizes and addresses transparency, and how then it translates this into its own transparency reporting and expectations for its member companies. This scoping paper is intended to guide a more detailed research paper that will examine in greater depth the specific themes and practices identified herein.

1 Transparency reporting is distinct from data access or data sharing.

2 Barbara Kowalczyk-Hoyer, "Transparency in Corporate Reporting: Assessing the World's Largest Companies," Transparency International, (2012): 37.

3 Joan Donovan, "Navigating the Tech Stack: When, Where and How Should We Moderate Content?," Centre for International Governance Innovation, October 28, 2019, https://www.cigionline.org/articles/navigating-tech-stack-when-where-and-how-should-we-moderate-content/.

4 Access Now, "Transparency Reporting Index - Access Now's Global Database," 2022, https://www.accessnow.org/transparency-reporting-index/; Ranking Digital Rights, "2020 Ranking Digital Rights Corporate Accountability Index," Accessed February 3, 2022, https://rankingdigitalrights.org/index2020/; Priya Kumar, "Ranking Digital Rights Findings on Transparency Reporting and Companies' Terms of Service Enforcement," Ranking Digital Rights, March 2016, https://rankingdigitalrights.org/wp-content/uploads/2016/03/RDR-Transparency-Findings.pdf.

Define objective(s)

The foundation of any good report requires clearly identifying the objective(s) of the information, the primary user(s) of the information, and the qualitative characteristics of useful information. Establishing the objective(s) of transparency reporting is the basis for effective reporting, and determining the audience(s) for such reports is a fundamental best practice that ensures the alignment of needs and expectations with what is covered by the report. Generally accepted government auditing standards, for example, require that the objectives of an audit be defined.[5]

Transparency reports can be used to "spark and grow the trust" of a company's user base,[6] signal to and inform policymakers, and force firms to build systems to enable them to capture and report on specific information. Other efforts are "narrowly designed to make government data more easily accessible to private sector and other stakeholders and do not attempt to consciously link these transparency mechanisms to accountability or participatory processes."[7] Defining objective(s) enables better design and assessment of whether transparency reporting is effective.

In voluntary transparency reporting, employing conceptual frameworks enables the development of standards that can be understood and correctly interpreted by all parties. Conceptual frameworks should be based on consistent concepts and the development of consistent reporting practices where no standard applies. Consider the following examples:

- The data-intensive Global Reporting Initiative (GRI) was designed to increase company transparency with respect to sustainability and improve decision making by companies and their stakeholders; and
- The Extractive Industries Transparency Initiatives (EITI), which implements the global standard for countries and companies to voluntarily disclose information on key aspects of the governance of oil, gas and mining revenues across the value chain, aims to encourage open and accountable management of those resources and "increase public and private sector responsiveness to citizen demands."[8] As a multi-stakeholder process operating at the country level, it provides data that can help inform reform efforts and strengthen public and corporate governance.

Good practice: Identify stakeholders and target audience(s)

Transparency reporting can be aimed at external audiences, but it also sends a signal to internal stakeholders about what is expected, acceptable, and important. There may be various audiences defined, but in most domains the general public or "users" are not typically an effective target audience. That said, in the case

5 "San Francisco Police Department Use-of-Force Data Audit: Interim Key Issue Report: Best Practices in Reporting Use-of-Force Data," City and County of San Francisco: Audits Division, City Services Auditor, Office of the Controller, December 18, 2019, https://sfgov.org/dpa/sites/default/files/SFPD_Key_Issue_Report_Use_of_Force_Data_Reporting_12_18_19_FINAL.pdf.

6 Peter Micek and Deniz Duru Aydin, "Non-Financial Disclosures in the Tech Sector: Furthering the Trend," in The Responsibilities of Online Service Providers, eds. Mariarosaria Taddeo and Luciano Floridi (Cham: Springer International Publishing, 2017), 241–61, https://doi.org/10.1007/978-3-319-47852-4_13.

7 Thomas Carothers and Saskia Brechenmacher, "Accountability, Transparency, Participation, and Inclusion: A New Development Consensus?," Carnegie Endowment for International Peace, October 20, 2014, https://carnegieendowment.org/2014/10/20/accountability-transparency-participation-and-inclusion-new-development-consensus-pub-56968.

8 Carothers and Brechenmacher, "Accountability, Transparency, Participation, and Inclusion."

of consumer-facing products and services that are used on a regular basis, communicating in a way that service users can understand transparency standards may be important. An audience may also be the industry itself, with one objective of such reporting to raise standards, inculcate norms, and/or promote best practices. Transparency reporting that requires extensive and intensive data collection, and which typically serves a specialist audience, would be a relevant issue to examine further as the objective(s) and audience(s) of transparency reporting are interdependent but also dependent on resources. The diversity of approaches is reflected in the following examples:

- The primary audience for GRI reports is specialists rather than the general public or consumers.[9] A primary audience for ESG reports is investors. In the case of ESG reports, there are 77 industry-specific SASB Standards, which are specifically aimed at helping businesses convey financial material sustainability information related to ESG issues to investors.
- The audience for financial transparency reporting is government regulators, as well as internal personnel and those charged with maintaining due diligence and observing regulations.
- The audience for the Internet Commission's report is primarily tech companies and regulators, mainly based in the UK and the Global North.[10]
- The audience for Ranking Digital Right's annual Corporate Accountability Index is its key stakeholders, namely participating tech firms and the digital rights community. However, it was designed with investors in mind and the index indicators were developed within the ESG framework (which is used by investors), with a focus on the social and governance. A few years in, the index is now being used by more and more investors, who turn to it to supplement and embed this data in their own activities.[11]
- The audience for technology sector transparency reporting on content moderation is not well defined.

√ Analyzing whether the stated objectives of reporting frameworks like GRI and EITI are in fact achieved and perceived as successful by key stakeholders and their intended audience can also inform whether ambitions for transparency reporting are realistic.[12]

√ Increased recognition that Global North/industrialized West companies have significant impact and influence on the Global South/developing countries underscores the need to consider various stakeholders and communities in disenfranchised localities. Similarly, specific communities affected by a company's operations, product, or services could be considered as a target audience.

What information should be in a transparency report?

What should GIFCT and/or its member companies include in transparency reports? On what issues is it reporting? What counts as data, how it is collected, presented, disseminated, and the costs involved are all relevant considerations for designing a transparency report or broader industry standards. A review of the literature and existing reporting across industries and sectors indicates that transparency reports should

9 See https://www.globalreporting.org/about-gri/; Stephanie Watts, "Corporate Social Responsibility Reporting Platforms: Enabling Transparency for Accountability," Information Technology and Management 16, no. 1 (March 1, 2015): 19–35, https://doi.org/10.1007/s10799-014-0192-2; Klaus Dingwerth and Margot Eichinger, "Tamed Transparency: How Information Disclosure under the Global Reporting Initiative Fails to Empower," Global Environmental Politics 10, no. 3 (August 1, 2010): 74–96, https://doi.org/10.1162/GLEP_a_00015.

10 Ioanna Noula, author interview with the Internet Commission, August 13, 2021.

11 Jan Rydzak and Amy Brouillette, author interview with Ranking Digital Rights, August 13, 2021.

12 Watts, "Corporate Social Responsibility Reporting Platforms"; Carothers and Brechenmacher, "Accountability, Transparency, Participation, and Inclusion."

typically include information on governance, policy, process, actions taken, impacts, results, and relations with government/authorities.

Good practices include providing quantitative and qualitative data and making sure that it is understandable,[13] including defining terms and abbreviations/relationships, explaining the methodology for collecting and analyzing the data, using data visualizations and examples, and explaining trends revealed by data.[14] Both ***data sources*** and ***presentation*** should include quantitative and qualitative information while contextualizing and explaining the information contained within.

A 2012 study of corporate transparency reporting by Transparency International recommended a series of baseline policies that all multi-nationals should adopt,[15] amounting to a set of best practices, some of which could be relevant for the GIFCT, including:

- Data transparency at the organizational, country, and corporate-level
- An informative website in at least one international language
- Including a list of all subsidiaries, affiliates, and related entities

Standardization and comparability

Standardization and consistency of data within a given industry or sector are widely recognized as a best practice. This enables comparison across entities, time, and data/issues. Taxonomies are important to allow data to be structured for sharing and comparison.[16] Verifiability and standardization of approaches for measuring are considered best practices across sectors. Minimum reporting requirements for an industry convey the basic expectations for transparency reports. Interoperability between different reporting requirements also helps "ensure that companies can collect information about performance on a given matter once and can use that same information to serve different objectives when the information is suitable for the needs of those different objectives."[17] Examples of such an approach include:

- The OECD Voluntary Transparency Reporting Framework (VTRF) version 1.0 was adopted in late 2021. It contains a set of baseline transparency questions for companies and a glossary of key terms that were developed and negotiated by a multi-stakeholder group of experts[18] over two years and then approved by the OECD's Committee on Digital Economy Policy. These minimum reporting standards on terrorist and violent extremist content (TVEC) moderation transparency represent a consensus from member countries

13 The provision of data in transparency reporting is distinct from the issue of access to data sources and the raw data itself.

14 As noted in several examples throughout. Also see, for example, Chloë Poynton, "Five Best Practices in Human Rights Reporting," BSR: Our Insights (blog), June 29, 2012, https://www.bsr.org/en/our-insights/blog-view/five-best-practices-in-human-rights-reporting.

15 It is notable that tech companies scored worst among the nine sectors analyzed in terms of transparency into corporate governance; see Kowalczyk-Hoyer, "Transparency in Corporate Reporting."

16 "Statement of Intent to Work Together Towards Comprehensive Corporate Reporting," Impact Management Project, World Economic Forum and Deloitte, September 2020, https://www.globalreporting.org/media/bixjk1ud/statement-of-intent-to-work-together-towards-comprehensive-corporate-reporting.pdf.

17 "Reporting on Enterprise Value Illustrated with a Prototype Climate-Related Financial Disclosure Standard," Impact Management Project, World Economic Forum and Deloitte, December 2020, https://29kjwb3armds2g3gi4lq2sx1-wpengine.netdna-ssl.com/wp-content/uploads/Reporting-on-enterprise-value_climate-prototype_Dec20.pdf.

18 The author was a member of the expert group.

about what the private sector should be regularly reporting on publicly,[19] and will be further elaborated and refined after a pilot period.[20] The OECD has created such reporting standards in other domains.

- A process is underway to develop SASB Standards for content governance to help investors assess the scope and scale of content moderation practices.[21]

Independent audits are an important principle across industries, and in some cases are embedded as requirements in mandatory transparency reporting.[22] Financial reporting is seen as the gold standard in transparency reporting because of its maturity and "adherence to internationally recognized accounting standards that bring transparency, accountability, and efficiency to financial markets around the world."[23] However, the cost of data collection, analysis, and verification can be significant, and thus pose significant hurdles for small companies and can even create an uneven playing field. For example, the U.S. Securities and Exchange Commission (SEC) conflict mineral rule requires certification and standardized audits by independent bodies. However, this can be prohibitively expensive for smaller companies, and there is a lack of oversight and verification of auditors by the SEC. Additionally, data access is often restricted to auditors or specific agencies. The broader issue of data access for researchers, journalists, or the public is a separate topic from transparency reporting, though it is relevant to issues of verifiability.[24]

Corporate Social Responsibility (CSR) and the Global Reporting Initiative (GRI)

Transparency is a core pillar of CSR. Companies that engage in CSR also report on those practices (typically voluntarily), although some jurisdictions mandate CSR reporting. An entire industry has emerged around CSR reporting, making it one of the most mature areas for further investigations into best practices.

With 95 percent of the world's biggest companies reporting on their CSR efforts, the vast majority of them use the GRI as the dominant reporting standard.[25] The GRI is data-intensive and provides standardization and comparability. It is also costly.

The GRI suggests four dimensions that would constitute a **best practice for transparent reporting: accuracy,**

19 Scott Morrison, "Media Release: More Action to Prevent Online Terror," Prime Minster of Australia, August 26, 2019, https://www.pm.gov.au/media/more-action-prevent-online-terror.

20 Jeremy West, "Why We Need More Transparency to Combat Terrorist and Violent Extremist Content Online," OECD Innovation Blog, September 15, 2020, https://oecd-innovation-blog.com/2020/09/15/terrorist-violent-extremist-content-internet-social-media-transparency-tvec/. The web portal is intended to be live in March 2022 at http://www.oecd-vtrf-pilot.org/

21 https://www.sasb.org/standards/process/active-projects/content-governance-in-the-internet-media-and-services-industry/.

22 One study found noteworthy "the importance of the national audit oversight bodies and the absence of professional bodies in the development of transparency reporting practice". Sakshi Girdhar and Kim K. Jeppesen, "Practice Variation in Big-4 Transparency Reports," Accounting, Auditing & Accountability Journal 31, no. 1 (January 1, 2018): 277, https://doi.org/10.1108/AAAJ-11-2015-2311; OECD, "OECD Best Practices for Budget Transparency," OECD Journal on Budgeting 1, no. 3 (May 16, 2002): 7–14, https://doi.org/10.1787/budget-v1-art14-en.

23 IFRS Foundation, "The case for global accounting standards," 2021, https://www.ifrs.org/use-around-the-world/why-global-accounting-standards/.

24 Policymakers are considering mandating access for independent and/or accredited researchers. See for example Article 30 and 31 of the EU's Digital Services Act and US proposals such as H.R. 3451 Social Media DATA Act. https://ec.europa.eu/info/sites/default/files/proposal_for_a_regulation_on_a_single_market_for_digital_services.pdf.

25 Watts, "Corporate Social Responsibility Reporting Platforms."

completeness, timeliness and relevance. GRI is built around six categories (Economic, Environmental, Human Rights, Society, Labor Practices and Decent Work, and Product Responsibility), each with its own subcategories and indicators. There are a total of 41 standards focused on an organization's outward impacts, which represent a broad consensus on good practice for reporting on a range of economic, environmental, and social impacts with respect to sustainable development.

The GRI offers three levels of reporting detail that firms can adhere to, which underscores the need for capacity/capability and intent when it comes to transparency. Such levels also are an acknowledgment of the fact that some firms may not have the capacity (technical, financial, etc.) to generate or collect particular data, which is a relevant concern for tech companies as well. The exploration of specific sectoral standards and a review of the Human Rights Assessment standard to identify principles, approaches, and good practices would help align technology platform reporting.

Determine level of analysis & granularity

An organization undertaking transparency reporting must determine what it means by 'organization' and whether a corporation will report data at the corporate or country level, and whether it covers subsidiaries, affiliates, and related entities. This should flow from its objective(s) and audience target(s).

Reporting on both subsidiaries *and* at the country level is a good practice in a range of sectors because it (a) enables evaluation of a company's activities and impact in each jurisdiction and by each service, which may cross jurisdictional boundaries and (b) "sheds light on any special arrangements between governments and companies, resulting in greater accountability."[26] This second aspect is particularly important for technology platforms and content moderation (and discussed further below). For example, a corporation like Facebook is a parent company with various subsidiaries (Instagram, WhatsApp) and regional offices, any or all of which could be the subject of its transparency reporting.

Some companies claim that reporting on a specific issue – for example anti-corruption efforts at the country level – would put them at a competitive disadvantage, whereas others view this as an internal aspect of risk management. The existence of the GIFCT as an industry-funded, membership non-profit organization that also oversees a technology (the shared hash and URL databases) to which various private ICT companies contribute material that would itself be the subject of their own transparency reporting raises the importance of defining the objectives in order to determine the level of analysis.

More granular approaches to transparency reporting can provide a better incentive to provide accurate information and thus result in better compliance. A review of tax transparency reporting in Europe found that country-by-country mandated reporting as opposed to regional level reporting resulted in fewer discrepancies and evasions.[27] Another European study found that greater transparency incentivized banks to improve their credit practices following a disclosure initiative introduced by the European Central Bank that required greater

26 It is notable that tech companies scored worst among the nine sectors analyzed in terms of transparency into corporate governance; see Kowalczyk-Hoyer, "Transparency in Corporate Reporting."

27 Niels Johannesen and Dan Thor Larsen, "The Power of Financial Transparency: An Event Study of Country-by-Country Reporting Standards," Economics Letters 145 (August 2016): 120–22, https://doi.org/10.1016/j.econlet.2016.05.029.

loan level information collection by banks and "stronger market discipline" in transparency reporting.[28] (See Tech Reporting Below for what this might mean in terms of transparency reporting on content moderation practices or impacts.)

√ What good practices would provide a roadmap for thinking about which levels of analysis to apply to GIFCT's transparency reporting? Should these apply to member companies? How would this contribute to setting standards for transparency in the tech sector?

Governance

Some companies report governance data in a stand-alone report, but in many cases it is incorporated into transparency reporting. If it is not in a stand-alone report, high-level data about governance as it relates to the objective(s) of the report and to policies and impacts assessed should be included in the transparency report itself (which can link to more extensive governance information).

While governance reporting is an entire field unto itself, good practices for providing basic information on governance should be addressed.[29] Furthermore, with Diversity, Equality, and Inclusion (DEI) becoming a more explicit goal for the private and public sectors alike, transparency reporting on governance and staffing should consider addressing this, particularly given concerns about censorship, definitions of problematic content, and geopolitical power dynamics inherent in the GIFCT's structure and mandate. The E.U.'s 2014 Accounting Directive added legally binding requirements for the disclosure of non-financial and diversity information by large companies and groups, in addition to existing financial disclosures, related to ESG issues.[30] Reports bound by its expectations must include information relating to "environmental matters, social and employee aspects, respect for human rights, anti-corruption and bribery issues, and diversity in their board of directors."

Addressing relationship with government

Several approaches, including CSR, ESG, and taxation reporting, provide information and details about the relationship of the organization or management with the government and/or relevant authorities (e.g. Ministry of Interior, the tax authority), suggesting that such reporting constitutes a good practice. Such governmental entities are also stakeholders for such reports.[31] For example, Oxfam's Behind the Brands Scorecard assessed the agricultural sourcing policies of the world's ten largest food and beverage companies, including transparency about those policies as well as corporate governance and influence efforts.[32]

Most content moderation-related transparency reports from tech firms have focused on government requests

28 Aytekin Ertan, Maria Loumioti, and Regina Wittenberg-Moerman, "Enhancing Loan Quality Through Transparency: Evidence from the European Central Bank Loan Level Reporting Initiative," Journal of Accounting Research 55, no. 4 (2017): 877–918.

29 "Corporate Governance: Simple, Practical Proposals for Better Reporting of Corporate Governance," Report Leadership, n.d., https://www.cimaglobal.com/Documents/Thought_leadership_docs/Governance/Report-Leadership-Corporate-Governance-Report.pdf; See also Ranking Digital Rights Corporate Transparency Index, https://rankingdigitalrights.org/index2020/explore-indicators.

30 See https://ec.europa.eu/info/business-economy-euro/company-reporting-and-auditing/company-reporting/corporate-sustainability-reporting_en.

31 Rachel F. Wang, Timothy C. Irwin, and Lewis K. Murara, "Trends in Fiscal Transparency: Evidence from a New Database of the Coverage of Fiscal Reporting," *Proceedings, Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association* 108 (2015): 1–37.

32 See https://www.behindthebrands.org/company-scorecard/.

of the company (see Transparency Reporting by tech companies). There is widespread agreement across stakeholder groups (private sector, government, civil society) that companies should report on government requests of ICT firms. In the 11 years since Google released the sector's first transparency report, more than 80 companies have started releasing transparency reports, indicating an emerging self-regulatory practice for companies to disclose privacy and freedom of expression threats (particularly with respect to government requests).[33] However, the mere existence of a transparency report does not mean that it is sufficient or fit for purpose, especially given that content moderation transparency reporting emerged from the tech industry in a seeming vacuum.

Transparency reporting on government requests should distinguish between content and non-content information; however, making such a determination about whether information is content or non-content is not always straightforward.[34] Further review and assessment of how specific data is classified under relevant U.S. and E.U. legal frameworks governing government access to communications would be helpful.

- The OECD released a report on good practices and minimum standards for reporting on government requests for access to personal data held by the private sector which should inform principles and practices for transparency reporting in the tech sector more broadly.[35]
- Similarly, EFF's annual assessment of content moderation policies, Who Has Your Back, focuses on identifying whether companies report on government takedown requests and provide meaningful notice and appeals process to users.[36] The nine years of reports are built around assessing industry best practices (from a human rights perspective) with respect to content moderation and government requests.

Notably for the tech sector, there is **little transparency reporting by government actors** on requests made to platforms either with respect to personal data, content moderation, or other purposes. The entire onus is currently put on ICT companies, although there is recognition and advocacy for governments to also produce such reporting, which would be in line with other frameworks.[37] Just as reporting by government and industry can reduce corruption and build trust through comparison and verifiability, it also has the potential to address concerns about undue political influence or pressure on tech companies outside of legitimate legal processes. If governments also reported on their moderation requests to platforms, then it would enable comparability and add a level of verifiability that currently does not exist.

Availability and access to data sources

Most third-party transparency reports, impact assessments, and rankings make use of publicly available data,

33 Whereas some information related to national security can only be reported in bands, companies and civil society alike have pushed to be allowed to provide greater detail. See Access Now, "Transparency Reporting Index."

34 Liz Woolery, Ryan Hal Budish, and Kevin Bankston, "The Transparency Reporting Toolkit: Best Practices for Reporting on U.S. Government Requests for User Information," 2016, https://dash.harvard.edu/handle/1/28552578.

35 José Tomás Llanos, "Transparency Reporting: Considerations for the Review of the Privacy Guidelines," OECD Digital Economy Papers, April 23, 2021, https://www.oecd-ilibrary.org/science-and-technology/transparency-reporting_e90c11b6-en.

36 See the 2015 edition: https://www.eff.org/who-has-your-back-government-data-requests-2015#best-practices.

37 The Australian eSafety Commissioner is required to publish annual reports, and Australia's new Online Safety Act, which took effect in 2022, contains some specific requirements as to what our future reports must contain.

including information found on websites, annual reports, and corporate transparency reports.[38] Embedded in these assessments and indices are ideas about what constitutes good practices when it comes to reporting transparently on corporate governance, company policies and practices, government requests, and user notification and remedy. Several ICT-related transparency reporting initiatives have emerged that further research should analyze to identify common principles, expectations, and practices with respect to data sources and access.[39]

Using publicly available data is a good practice because it enables verification and comparison, but often raw data emerges from proprietary business operations, so a better understanding of how data is generated and verified in other sectors is needed to inform the technology sector's approach. Some reporting is required by law and/or mandates disclosure of specific information. The U.S. SEC and OECD, for example, have reporting requirements for companies involved in conflict minerals that require disclosure about their uses and information about their due diligence processes.

- The OECD process has been deemed to be the most established one for compliance with conflict minerals reporting and merits closer review. Article 40 of the E.U.'s Eighth Directive, which sets out transparency reporting requirements for accounting and audit firms, is quite broad and leaves some discretion to member states in relation to the implementation of the directive into local law.[40]
- The Global Network Initiative (GNI) human rights assessment and the Internet Commission's responsibility evaluation framework both rely on confidential information obtained from companies under review, including data and interviews, and require researchers/auditors to sign non-disclosure agreements (NDAs).[41] This limits the replicability or verifiability of these assessments, but provides access to proprietary information and officials that would otherwise not be available.
- Many researchers, civil society groups, and journalists dislike and will not sign NDAs, and these are increasingly seen as problematic.[42]

Dissemination

A critical part of a successful transparency report is ensuring that it gets to the people who need it. Using the correct, publicly available technology for access to reporting information is important.[43] A best practice from various domains is to make data contained within transparency reports publicly available in a machine-

38 This is due to a range of factors including access to relevant data, time and costs required to collect or generate data, and a commitment to transparency of their own processes.

39 PCIO Baseline Datasets, "Transparency Reporting & Data Sharing," Partnership for Countering Influence Operations, accessed February 18, 2022, https://ceip.knack.com/pcio-baseline-datasets#transparency--data-sharing/?view_69_page=1&view_69_filters=%7B%22match%22%3A%22and%22%2C%22rules%22%3A%5B%7B%22field%22%3A%22field_448%22%2C%22operator%22%3A%22contains%22%2C%22value%22%3A%22government%22%2C%22field_name%22%3A%22Stakeholder%20%22%7D%5D%7D.

40 S. Girdhar and K.K. Jeppesen, "Practice variation in Big-4 transparency reports," *Accounting, Auditing & Accountability Journal* 31, no. 1 (2018): 261–285, https://doi.org/10.1108/AAAJ-11-2015-2311.

41 The author was a board member of the GNI; Noula, author interview.

42 Based on the author's decade of experience working in the sector and specific discussions about the use of NDAs. For example, the author previously worked for the Committee to Protect Journalists which refused on principle to sign NDAs with technology platforms. This issue was also discussed in civil society groups including the Christchurch Call Advisory Network.

43 "Statement of Intent," Impact Management Project; Renata Avila et al., "Global Mapping of Technology for Transparency and Accountability," *Transparency Accountability Initiative*, 2010, https://www.transparency-initiative.org/wp-content/uploads/2017/03/global_mapping_of_technology_final1.pdf.

readable format, including explanatory charts, infographics, etc., that tell the story embedded in the data. Ensuring they are presented in a way that addresses the target audience/s is also an important factor.

- Embedded in the EITI is a commitment to open data, accessibility, and compatibility, making country-level data available through an **API** and via direct downloadable files along with the individual annual reports.
- Further research into how to best disseminate reports is needed, taking into consideration the objective(s) and audience(s).[44]

Lesson: Content moderation

Transparency reporting in the tech sector is overwhelmingly focused on content moderation issues, though how these are addressed and assessed has evolved over the past decade. Access Now's Transparency Reporting Index collects links to transparency reports from major internet and telecom companies around the world by year, but does not assess the quality of those reports.[45] That said, common across all reports is reporting on user privacy data, specifically the number of requests for user data coming from governments, police, or other law enforcement agencies (and in some cases intelligence agencies).[46] This can mean distinguishing between criminal and national security requests, although the level of granularity varies widely, with some companies specifying the specific category of content while others lump them together.

Just as the 2013 Snowden revelations about NSA spying and access to U.S. tech companies' data propelled more ICT entities to report on government access and removal requests, the cumulative efforts to counter violent extremism and terrorism online appears to have propelled an increasing number of companies to engage in reporting on content moderation as it relates to terrorism/CVE.[47] Similarly, since 2017 and amid the COVID-19 "Infodemic," reporting on information/influence operations has become a more regular practice among the largest social media firms, which are most often the target of such campaigns.[48] Facebook, Google/YouTube, and Twitter began reporting on these platform takedowns campaigns in 2017 but have significantly increased since then. However, there is no shared definition of misinformation or other problematic content categories, making comparison difficult. These reports are ad hoc and report on dimensions defined by the company at hand, sometimes in collaboration with civil society groups that identified the information operation.[49]

A comprehensive 2016 survey of U.S. internet and telecom transparency reporting identified eight principles and illustrates the importance of including key metadata (e.g. date) and a static URL for each report.[50] This study identified clear and granular categorization of specific legal processes as well as reporting on the

44 Avila et al., "Global Mapping of Technology."

45 Access Now, "Transparency Reporting Index."

46 Micek and Aydin, "Non-Financial Disclosures in the Tech Sector."

47 OECD, "Transparency Reporting on Terrorist and Violent Extremist Content Online: An Update on the Global Top 50 Content Sharing Services," OECD Digital Economy Papers, accessed August 12, 2021, https://www.oecd.org/digital/transparency-reporting-on-terrorist-and-violent-extremist-content-online-8af4ab29-en.htm.

48 "Disinfodex," Partnership for Countering Influence Operations, accessed February 5, 2022, https://disinfodex.org/; Jon Bateman Smith and Victoria Natalie Thompson, "How Social Media Platforms' Community Standards Address Influence Operations," Carnegie Endowment for International Peace, April 1, 2021, https://carnegieendowment.org/2021/04/01/how-social-media-platforms-community-standards-address-influence-operations-pub-84201.

49 "Disinfodex," Partnership for Countering Influence Operations.

50 Woolery, Budish, and Bankston, "The Transparency Reporting Toolkit."

subjects of requests and how users are impacted; comprehensive explanations of legal processes; and the need for standardization of definitions in order to achieve standardization in categories as best practices. It also recommends including a detailed and illustrative (though non-exhaustive) list of how a provider can respond along with provider-specific examples in various categories and definitions. But the framework also suggests that in addition to best practices, good, standard, and notable practices may also provide useful information about transparency reporting principles and implementation.

There appears to be a trend toward providing greater granularity and reporting on a company's own content moderation policies and aggregate impacts on specific types of content (and not just at the behest of government/law enforcement). This could indicate an emerging best practice toward specificity and comprehensiveness with respect to content and account removals, with some advocating for greater details about content and account enforcement, actions taken, and rationale.[51] Amid ongoing advocacy from civil society and academia, policymakers in the E.U. and the U.S. appear poised to reinforce this emergent trend through mandates or regulation.[52] An emerging consensus on minimum expectations for reporting on content moderation by ICTs has emerged through various reports, recommendations, frameworks, and principles specific to the tech industry.[53] Further research should examine what expectations are embedded in such proposals, as they reflect normative expectations of key stakeholders and may indicate areas of consensus.[54]

For example, the Internet Commission's transparency reporting framework for social media content moderation proposes five categories to be assessed with qualitative and quantitative indicators: reporting, moderation, notice, process of appeal, resources, and governance.[55] These were built in part on the Santa Clara Principles on Transparency and Accountability in Content Moderation,[56] which spell out a set of minimum expectations for reporting, and are in line with the terms of service recommendations from the Internet Governance Forum Dynamic Coalition on Platform Regulation.[57] The 2.0 Principles outline general expectations for a broader range of content moderation actions and policies. The Global Disinformation Index recommends tech companies use unique error codes corresponding to the policy under which a piece of content was removed and developing a common notice and takedown regime, both of which would contribute to reporting standardization and comparison.[58]

One overarching challenge with the tech sector as compared to others is that policies and standards governing

51 Based on the author's observations and engagement in a variety of venues where tech sector transparency reporting is being discussed.

52 See the EU's Digital Services Act and the US Senate's proposed Algorithmic Justice and Online Platform Transparency Act (S. 1896).

53 Spandana Singh and Leila Doty, "The Transparency Report TrackingTool: How Internet Platforms Are Reporting on the Enforcement of Their Content Rules," New America: Open Technology Institute, December 9, 2021, http://newamerica.org/oti/reports/transparency-report-tracking-tool/; "Transparency Reporting & Data Sharing," Partnership for Countering Influence Operations.

54 As the 2021 OECD TVEC report notes, "the number of jurisdictions that have TVEC-related laws and regulations in force or under consideration is growing, but they are not consistent, either. That presents a risk of divergent reporting standards and requirements." It also poses an opportunity to identify commonalities and minimums.

55 See https://www.dropbox.com/s/fdzvwqeyosdezb9/The%20Internet%20Commission%20%E2%80%93%20transparency%20reporting%20framework.pdf?dl=0.

56 See https://santaclaraprinciples.org/.

57 See Annex 12.5.1 Degree of Monitoring in Luca Belli et al., "Platform Regulations: How Platforms Are Regulated and How They Regulate Us," *FGV Direito Rio*, 2017, https://bibliotecadigital.fgv.br/dspace/handle/10438/19402.

58 Benjamin T. Decker and Tim Boucher, "Disrupting Online Harms: A New Approach," The Global Disinformation Index, July 2021, https://disinformationindex.org/wp-content/uploads/2021/07/2021-07-23-Disrupting-Online-Harms-A-New-Approach.pdf.

content and behavior on a specific platform are regularly revised based on new or emerging issues. This means that there is no required minimum or standard set of reporting criteria and that the internal criteria shift. Because there is typically no archive of previous policy iterations, it is difficult to assess how these evolve over time, much less to audit consistency with internal guidelines.
A more detailed review and analysis focused on multi-stakeholder and consensus-driven frameworks should be conducted to identify good practices, minimum transparency expectations, and technical feasibility for reporting on content moderation.

Beyond content moderation: Digital responsibility

Ranking Digital Rights' Corporate Accountability Index evaluates the policies and practices of digital and telecom companies that affect human rights on an annual basis using publicly available information.[59] Companies are evaluated on a range of indicators that fall broadly into three buckets: governance, freedom of expression, and privacy. Each one contains a range of indicators related broadly to access to various types of information, including several indicators related to content and account moderation, as well as remedy and appeals and other related processes. These criteria mirror much of the same type of information that is available in tech company transparency reports.

The Internet Commission's Evaluation Framework for Digital Responsibility proposes a detailed set of qualitative and quantitative indicators related to organization, people, governance, content moderation, automation, and safety. It uses public data as well as proprietary information and interviews.[60]

Good practices

The following summary of good practices apply to data and transparency and impact reporting in a range of sectors and across approaches:

- User-friendly and concise
- Accurate and Clear
  - Use of illustrative examples
  - Use of tables, charts, infographics
  - Complete
  - Glossary as needed
  - Explains trends or data interpretation to avoid misinterpretation
  - Considers which data may need additional context or interpretation
- Accessible
  - Downloadable
  - Machine readable/API
- Timely
- Relevant

59 Since policies and URLs change and information may be buried in a website, RDR maintains a database of snapshots of the corresponding content for each indicator and an explanation of its assessment, meaning that its assessment is both verifiable and replicable.

60 See https://drive.google.com/file/d/13aaItNDoynvXHeNZLF2yjIliQsRcdbvb/view.

- Verifiable
  - Replicable/Auditable
  - Access to data/information sources

Figure 1: Examples of how an assessment of how San Francisco's police use-of-force transparency reports measured up to best practices[61]

Exhibit 1: The Police Department Can Improve Its Use-of-Force Reports by Aligning Them With Best Practices

Best Practices for Reporting Data		EIS Report	96A Report
Context	Reports should provide context to assist users in interpreting data and facilitate informed decision making.	✗	−
User needs	Reports should include data that is summarized, stratified, and provided in appropriate detail to meet the needs of stakeholders relying on the data.	✗	✓
Key points	Reports should include a concise and organized executive summary to improve the structure of the report and ensure users can easily follow relevant points.	✗	−
Visualization	Reports should represent data, especially more complex data, through graphics that accurately show trends, relationships, and the most significant information.	✗	−
Open Data	Data that supports reports should be available to increase public trust.	✗	✗
Accuracy and completeness	Stakeholders should be able to rely on the accuracy and completeness of the data *underlying* reports to make informed decisions.	CSA will assess this in its full audit*	

✓ Complies with best practice − Partly complies with best practice ✗ Does not comply with best practice

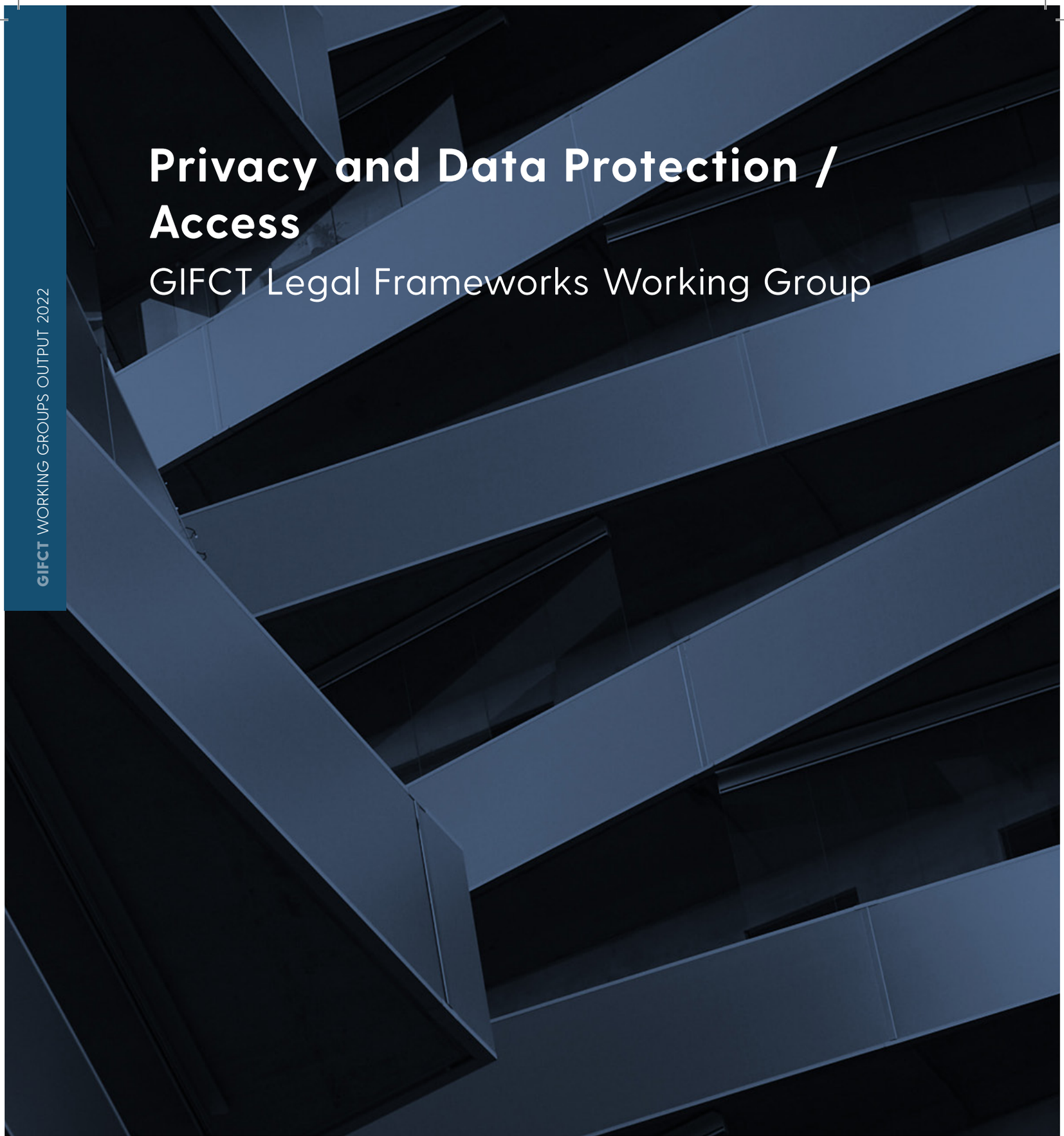
*As part of its full audit, CSA is assessing the accuracy and completeness of the data underlying the EIS and 96A reports.
Source: Best practices from publications on writing statistics for governments; compliance with best practices assessed by CSA.

Next steps:

The references and organizations identified in this scoping paper provide a jumping-off point for identifying who to interview[62] and what reports to analyze. There are several databases and existing reports on content moderation and tech sector transparency that can form the basis for further study. Although there are published assessments and analyses of standard and best practices, there are many aspects of the process that can only be gleaned through interviews and analysis of specific reports themselves. These include the collection and analysis of data, the costs involved in different approaches and good practices and how tradeoffs between good practices and resource considerations are made, the internal mechanics in terms of staffing and process that goes into the production of the report, and what has worked for effective dissemination of the report. Interviews can provide further insight into how an organization decided what and when to measure and how it negotiated among its various stakeholders and the expectations of the field in which they are situated. Assessing whether a given set of transparency reporting objective(s) were achieved and whether they reached their target audience(s) – and what worked or didn't in terms of communicating information in a meaningful way (e.g., traffic to relevant webpages, efficacy of printed reports, and adherence to other best practices laid out above) – will be an important part of the research.

61 Screenshot from https://sfgov.org/dpa/sites/default/files/SFPD_Key_Issue_Report_Use_of_Force_Data_Reporting_12_18_19_FINAL.pdf.

62 This should include a range of stakeholders including government and civil society actors.



GIFCT WORKING GROUPS OUTPUT 2022

Privacy and Data Protection / Access

GIFCT Legal Frameworks Working Group

GIFCT
Global Internet Forum to Counter Terrorism

Dia Kayyali
Mnemonic

Introduction

Last year's Legal Frameworks Working Group whitepaper focused on "the issues relating to the work of technology companies disrupting terrorist and violent extremist content (TVEC) and the intersection with access to data" by parties such as "industry, research, academic, or civil society actors."[1] It identified high-level issues associated with this topic and provided some recommendations. This paper dives deeper into the topic through case studies, further legal and literature reviews, and interviews.

Online platforms, in particular social media platforms, host a variety of content that they might, for one reason or another, determined to be "terrorist or violent extremist content." The process of designating content as TVEC can be accelerated through participation in GIFCT, both through access to the hash database and information sharing. This content and associated data, whether or not it actually violates any laws or platforms' rules, is valuable for a variety of important purposes. Those purposes include journalism, research on content moderation practices, industry research (for example, small platforms learning about content moderation), misinformation/disinformation research, probes into specific violent attacks with online components, and finally open-source investigations into human rights violations. However, providing free and open access to this data is a clear non-starter due to privacy and security issues.

Each of these uses of data presents unique challenges. This paper was initially meant to be very broad but it became clear through research that there is no "one size fits all" solution. In the process of this research, the re-escalation of hostilities against Ukraine started, emphasizing the importance of open-source investigations and the susceptibility of this content to removal. Furthermore, a GIFCT Content Incident Protocol was activated on May 15 after a perpetrator livestreamed himself at a grocery store carrying out a mass shooting that targeted Black Americans.[2]

Thus, this paper focuses on the use of content hosted online for open-source investigations. The paper also outlines broad concerns regarding data preservation and access and touches on some of the challenges posed by government requests for data associated with offline attacks with an online component.

The Legal Frameworks Working Group recommends that civil society organizations and government work together to find a legislative solution that would provide a legal avenue for the International Criminal Court and United Nations investigative bodies to access removed content in a way that would not require platforms to violate the Stored Communications Act (SCA) or the European Union General Data Protection Regulation. It also recommends increased transparency from platforms and further research into access for other purposes. Finally, the working group urgently recommends that GIFCT commission (or work with governments and civil society organizations to produce) research into the kinds of data governments need to understand terrorist and violent extremist use of the Internet after attacks. Between the time this paper is written and the time it is published, the Content Incident Protocol debrief may provide answers to some of these questions, but regardless, this final recommendation should be a priority specifically for GIFCT as the most relevant forum for discussions about

1 Legal Frameworks Working Group, "Legal Frameworks Report," Global Internet Forum to Counter Terrorism, July, 2021, https://gifct.org/wp-content/uploads/2021/07/GIFCT-LegalFrameworks-WGroup.pdf.

2 Global Internet Forum to Counter Terrorism, "Update: Content Incident Protocol Activated in Response to Shooting in Buffalo, New York United States," May 17, 2022 https://gifct.org/2022/05/14/cip-activated-buffalo-new-york-shooting/.

online components of offline attacks.

Open-source investigations

Organizations like Bellingcat and Syrian Archive do open-source investigations (OSI) into human rights violations. OSI rely on user-generated content and are often key in investigating violations in places where traditional investigations may not be possible for a variety of reasons. Organizations discover content through various methods, archive it, and verify that it actually depicts what it says it does, putting together public or private data sets to aid in investigating specific cases. This content is particularly susceptible to being removed (whether correctly or not) by content moderation practices for a variety of reasons. The content can be very graphic and related to groups that are designated as terrorist organizations by U.S. or UN lists, by platforms themselves, or by other governments. Furthermore, considerable amounts of human rights related content is not in English but instead in non-Latin languages such as Arabic, Burmese, and Ukrainian. These factors in content removal have been explored by myriad content moderation and natural language processing experts and are not the focus of this paper, although GIFCT member companies would benefit from more focused research on these topics.

Ultimately, data that could be used by the International Criminal Court or other bodies may be lost and destroyed forever. This is especially challenging because it may be the only evidence available. However, asking platforms to preserve this content, or to provide access beyond their existing procedures for law enforcement, raises many legal and ethical issues.

The Berkeley Protocol on Digital Open Source Investigations, co-published by the United Nations and the Human Rights Center at the University of California, Berkeley, School of Law provides indispensable guidance for using this content, including consideration of ethical and legal issues.[3]

The Mnemonic Experience

In July of 2017, the NGO Syrian Archive and myriad Syrian human rights defenders that had amassed vast collections of documentation from Syria noticed that their content was being deleted and accounts being suspended more rapidly than ever before on YouTube.[4] These mass removals started less than a month after Google announced it would now be using machine learning to detect content it deemed to be "terrorist and violent extremist content."[5]
Many news outlets reported on the issue, and Syrian Archive started to track removals from their own collection, as well as assist account owners whose content had been removed. YouTube voluntarily restored thousands of videos over the next several years as Syrian Archive, along with the NGO WITNESS, worked to address the causes of these removals.[6]

3 United Nations Office of the High Commissioner and the Human Rights Center at the University of California Berkeley School of Law, "Berkeley Protocol on Digital Open Source Investigations," HR/PUB/20/2, 2022, https://www.ohchr.org/sites/default/files/2022-04/OHCHR_BerkeleyProtocol.pdf.

4 Raja Althabani and Dia Kayyali, "Vital Human Rights Evidence in Syria is Disappearing from YouTube," WITNESS blog, August 30, 2017, https://blog.witness.org/2017/08/vital-human-rights-evidence-syria-disappearing-youtube/.

5 The YouTube Team, "An update on our commitment to fight violent extremist content online," YouTube Official Blog, October 17, 2017, https://blog.youtube/news-and-events/an-update-on-our-commitment-to-fight/.

6 Hadi Al Khatib and Dia Kayyali, "YouTube is Erasing History," New York Times, October 23, 2019, https://www.nytimes.com/2019/10/23/opinion/syria-youtube-content-moderation.html.

The use of automation to detect and remove content, including the GIFCT hash database, has undisputedly continued to lead to high rates of removal. Statistics and statements from Meta, YouTube, and Twitter from the last five years confirm this.[7]

Mnemonic is the parent organization for Syrian Archive. Its "lost and found" project has tracked content removals to the greatest extent possible by comparing archived collections of content against the original URLs of that content and determining what is still online. For example, currently about a quarter of Syrian Archive's collection gathered from YouTube is no longer online. Some platforms actually provide enough information to determine the cause of removal while some do not. Mnemonic also relies on insight from relationships with users and organizations that post content and share why that content is removed.

Mnemonic sometimes sees direct spikes in removal in relation to external events. In Yemen, for example, Mnemonic found a large and unexplainable jump in content removals of Houthi accounts after the Trump administration indicated that it would be designating Houthis as a "foreign terrorist organization" (FTO).[8] From October 1 to November 30, 2020, "53,846 tweets became unavailable. That's highly abnormal. By comparison, it's more common for [Yemeni Archive] to see a few thousand tweets become available per month." These accounts were important sources of human rights evidence. Twitter said that these removals were due to its "Platform Manipulation and Spam" policy, but the timing of the removals remains suspicious. In Palestine, content removals often skyrocket during periods of intensified IDF activity. The Palestinian organization 7amleh has received reports of removals from users, and Human Rights Watch has also tracked extensive removals from their own investigations.[9]

What kind of data and how is it used?

Content hosted by online platforms and associated data, in particular social media platforms, is increasingly important for international and domestic justice mechanisms. As noted above, investigators discover user-generated content and associated data, such as the identity of users posting the content, through OSI.

The use of online content is not always readily apparent from outside of justice mechanisms. It can be used directly as evidence when it can meet applicable legal standards. However, it can also be used in a variety of other ways. It can be very helpful in early stages, where it can inform case building – i.e., help investigators know where to focus. It can also help courts and other international mechanisms create operational security and witness protection plans, locate witnesses and corroborate their statements, and establish timelines. This is

7 Monika Bickert and Brian Fishman, "Hard Questions: What Are We Doing to Stay Ahead of Terrorists?," Meta Newsroom, November 18, 2018, https://about.fb.com/news/2018/11/staying-ahead-of-terrorists/; Erin Saltman, " Identifying and Removing Terrorist Content Online: Cross-Platform Solutions," The Raisina Edit (blog series), 2022, https://www.orfonline.org/expert-speak/identifying-and-removing-terrorist-content-online/; Zoe Strozewski, "Twitter Suspended 44K Accounts for Promoting Terrorism, Violent Orgs in First Half of 2021," Newsweek, January 25, 2022, https://www.newsweek.com/twitter-suspended-44k-accounts-promoting-terrorism-violent-orgs-first-half-2021-1672868.

8 Dia Kayyali, "What happens when the US decides to designate a group as a terrorist organization? Insights from Mnemonic," Mnemonic, February 18, 2022, https://mnemonic.org/en/content-moderation/What-happens-terrorist-designation.

9 7amleh- The Arab Center for the Advancement of Social Media, "The Attacks on Palestinian Digital Rights," May 21, 2021, https://7amleh.org/2021/05/21/7amleh-issues-report-documenting-the-attacks-on-palestinian-digital-rights; Belkis Wille, "'Video Unavailable', Social Media Platforms Remove Evidence of War Crimes," Human Rights Watch, September 10, 2020, https://www.hrw.org/report/2020/09/10/video-unavailable/social-media-platforms-remove-evidence-war-crimes.

important, because these uses of data are not readily visible to stakeholders outside of these mechanisms, and in fact these mechanisms cannot always be forthcoming about their uses of data for security or other reasons.

Last year's paper identified the following kinds of data which this paper refines. As noted last year, these categories are not mutually exclusive:

- Personally identifiable information (PII): this would include things like people's legal names, images of their faces, recordings of someone's voice, or addresses.
- TVEC: It should be noted that "TVEC" is a disputed term because it relies on the varied definitions of terrorism and violent extremism used by different platforms and governments. It's also not completely clear what part of the content "TVEC" refers to. Instead, the term "user-generated content" is more helpful. This term refers to the actual content itself, i.e., the words of a post or the video file or image file of a post, as well as associated data such as the title of a video. This would also include comments posted by other users.
- Metadata relating to content: This would be information like the IP address a particular piece of content was posted from.
- Non-personal descriptive data: This includes hashtags, key phrases, titles of broadcasts, etc. For purposes of this paper, this category is subsumed under "user-generated content."
- Contextual information about the sources of data.

As noted, these various categories of data can be used in different ways, and they implicate different ethical and legal concerns. For purposes of this paper, "data" is equivalent to "user-generated content" when referring to open source investigations, and is the main focus. The United Nations and the International Criminal Court, cases for accountability in Syria, and the growing number of justice efforts for war crimes in Ukraine provide helpful examples of the use of this data.

United Nations

The United Nations has a long history of investigating human rights violations through special investigatory bodies designed for that purpose, but only with the Syrian conflict did OSI start providing meaningful insight. The International, Impartial, and Independent Mechanism on Syria (IIIM) was established by the UN General Assembly (UNGA) in 2016 after vetoes in the UN Security Council prevented the referral of the Syrian situation to the International Criminal Court (ICC). The majority of evidence preserved by the IIIM is digital, and the IIIM works closely with civil society organizations to collect that evidence.

Similarly, the UN established the Independent Investigative Mechanism on Myanmar (IIMM) in 2018. Like the IIIM, the IIMM has relied heavily on digital evidence. Unlike the IIIM, the IIMM has made its struggle with social media content very public, excoriating Meta in particular for its role in the genocide of the Rohingya and calling on Meta to preserve evidence. According to the UN's IIMM's first report, Facebook was found to be "the internet" in Myanmar, with Myanmar officials being able to spread anti-Muslim and anti-Rohingya hate speech and disinformation.[10] Meta has voluntarily complied with some of the IIMM's requests, but the struggle to obtain

10 United Nations Human Rights Council, "Report of Independent International Fact-Finding Mission on Myanmar," A/HRC/39/64, September 12, 2018, https://www.ohchr.org/en/hr-bodies/hrc/myanmar-ffm/reportofthe-myanmar-ffm.

further evidence has been well documented in The Republic of the Gambia v. Facebook, Inc.[11]

There have been several interesting developments for the UN. First, France recently passed legislation enabling "information to be transmitted from French courts" to the IIIM.[12] This could set a precedent for other national governments, including the United States. Second, legal experts from Oxford just released a report recommending that the UN provide permanent support for UN-mandated investigations, and presents two options: "the establishment of a standing, independent UN investigative support mechanism (ISM)" or "the establishment of a permanent investigative support division (ISD) within the Office of the High Commissioner for Human Rights (OHCHR)."[13] As noted in the recommendations section of this paper, either of these options would make it easier to craft due process respecting transfers of evidence from social media platforms to the UN. **If the standing mechanism were to become reality, it could make it much easier to ensure data is collected and stored in a timely, legal, and ethical manner that protects the safety and security of people whose PII is included in that data.**

Cases for accountability in Syria

As noted above, the use of data from online sources is not always readily apparent, but it has been a part of every case for accountability in Syria. In a historical first, in January 2022, the Koblenz Higher Regional Court in Germany convicted a senior Assad government official for crimes against humanity.[14] However, this was not the first Syria-related case. Syrian Archive has worked with other NGOs to file cases in several courts in France, Sweden, and Germany about the use of chemical weapons by the Syrian government, seeking to hold the government accountable.[15] Syrian Archive has been a party to every one of these cases, and the organization's extensive open-source chemical weapons database has provided indispensable evidence for each one.[16] Even before these cases were filed, evidence from Syrian Archive was successfully used in a 2019 prosecution of three Belgian firms that provided a sarin gas precursor to the government.[17] Open-source evidence has also been used in Swedish and German courts in several war crimes trials.

11 The Republic of Gam. v. Facebook, Inc., Civil Action 20-mc-36-JEB-ZMF (D.D.C. September 22, 2021).

12 The Ministry for Europe and Foreign Affairs (France), "Jurisdiction of French courts over crimes against humanity," February 9, 2022, https://www.diplomatie.gouv.fr/en/french-foreign-policy/international-justice/news/article/jurisdiction-of-french-courts-over-crimes-against-humanity-9-feb-2022.

13 Federica D'Alessandra et al., "Anchoring Accountability for Mass Atrocities: The Permanent Support Needed to Fulfil UN Investigative Mandates," The Oxford Institute for Ethics, Law, and Armed Conflict, May, 2022, https://www.bsg.ox.ac.uk/sites/default/files/2022-05/Anchoring%20Accountability%20for%20Mass%20Atrocities%20Report.pdf.

14 European Center for Constitutional and Human Rights, "First criminal trial worldwide on torture in Syria before a German court," June 2, 2022, https://www.ecchr.eu/en/case/first-criminal-trial-worldwide-on-torture-in-syria-before-a-german-court/

15 Syrian Center for Media and Freedom of Expression, "Survivors and the Syrian Center for Media and Freedom of Expression, with Support from Syrian Archive and the Justice initiative, Seek French Criminal investigations of Chemical Weapons Attacks in Syria," March 2, 2021, https://scm.bz/en/scm-statements/survivors-and-the-syrian-center-for-media-and-freedom-of-expression-with-support-from-syrian-archive-and-the-justice-initiative-seek-french-criminal-investigations-of-chemical-attacks-in-syria; Simon Johnson, "Victims of chemical attacks in Syria file complaint with Swedish police," Reuters, April 19, 2021, https://www.reuters.com/world/middle-east/victims-chemical-attacks-syria-file-complaint-with-swedish-police-2021-04-19/.

16 Syrian Archive, "Chemical Weapons Database," May 18, 2022, https://syrianarchive.org/en/datasets/chemical-weapons.

17 Syrian Archive, "Antwerp court convicts three Flemish firms for shipping 168 tonnes of isopropanol to Syria," February 7, 2019, https://syrianarchive.org/en/investigations/BI-sentencing; Jeff Deutch and Kristof Clerix, "Belgium illegally shipped 168 tonnes of sarin precursor to Syria," Syrian Archive, 2018, https://syrianarchive.org/en/investigations/belgium-isopropanol.

Ukraine

Like Syria, the escalation of Russian aggression toward Ukraine is being thoroughly documented on social media platforms. Unlike Syria, however, there has been a very rapid response from governments and international bodies interested in prosecuting war crimes, and an immediate acknowledgment of the importance of digital evidence. Civil society organizations have been able to take lessons learned about finding and archiving content – particularly from Syria, where Russian military and equipment carried out many attacks on civilian infrastructure – and apply those lessons to Ukraine. The civil society response has been almost immediate. Ukraine provides a clear example of the importance of content, as well as the dangers of deletion of that content.

There are myriad overlapping investigations, including an ICC investigation, a UN Commission of Inquiry, and investigations by the Ukrainian General Prosecutor.[18] In fact, the first trial for a war crimes case in Ukraine concluded after the Russian soldier on trial pled guilty.[19] The European Commission is also supporting ICC and creating own mechanism for storing content, and jurisdictions, and Germany's Federal Prosecutor has also started an investigation into war crimes.[20] Finally, on May 17, the United States Department of State announced the establishment of the Conflict Observatory, a program that will document, verify, preserve, analyze and share open-source evidence: "publicly and commercially available information, including satellite imagery and information shared via social media."[21] Although the Conflict Observatory was created specifically to respond to Russian war crimes in Ukraine, it may provide a model for other conflicts.

Mnemonic and Bellingcat are two of the leading organizations working on archiving and verifying Ukrainian content in collaboration with Ukrainian organizations (including the 5 am coalition), and they have observed several trends.[22] First, content from Ukraine has been removed at a lower rate than Syria, for unfortunately obvious reasons – few of the participants in the conflict are designated terrorist organizations, content is not being posted in Arabic, and there is wide public and governmental support for Ukrainians. The presence of the Azov battalion, which is on Meta's DIO list, has had an impact on content. Meta had an exception for

18 International Criminal Court, "Ukraine: Situation in Ukraine," March 2, 2022, https://www.icc-cpi.int/ukraine; United Nations Human Rights Council, Forty-ninth session, "Situation of human rights in Ukraine stemming from the Russian aggression," A/HRC/RES/49/1, March 7, 2022, https://documents-dds-ny.un.org/doc/UNDOC/GEN/G22/277/44/PDF/G2227744.pdf?OpenElement; Greg Myre, "Ukraine begins prosecuting Russians for war crimes," NPR, May 14, 2022, https://www.npr.org/2022/05/14/1098941080/ukraine-begins-prosecuting-russians-for-war-crimes.

19 Associated Press, "Russian soldier pleads guilty at Ukraine war crimes trial," May 18, 2022, https://apnews.com/article/russia-ukraine-kyiv-moscow-war-crimes-61c89e6c73541f3fa2364dde1498df72.

20 Bokan Pancevski, "Germany Opens Investigation Into Suspected Russian War Crimes in Ukraine," The Wall Street Journal, March 8, 2022, https://www.wsj.com/livecoverage/russia-ukraine-latest-news-2022-03-08/card/germany-opens-investigation-into-suspected-russian-war-crimes-in-ukraine-bNCphal-WE30f2REH8BCi; Directorate-General for Neighbourhood and Enlargement Negotiations, "Russian war crimes in Ukraine: Commission welcomes European Parliament's adoption of Eurojust's reinforced mandate," European Commission, May 19, 2022, https://ec.europa.eu/neighbourhood-enlargement/news/russian-war-crimes-ukraine-commission-welcomes-european-parliaments-adoption-eurojusts-reinforced-2022-05-19_en.

21 Office of the Spokesperson, "Promoting Accountability for War Crimes and Other Atrocities in Ukraine," United States Department of State, May 17, 2022, https://www.state.gov/promoting-accountability-for-war-crimes-and-other-atrocities-in-ukraine/.

22 Alfred Landecker Foundation, "Mnemonic - Ukrainian Archive, Preservation, verification, and investigation of open-source documentation concerning human rights violations in Ukraine," May 19, 2022, https://www.alfredlandecker.org/en/article/introducing-mnemonic; Human Rights Centre ZMINA, "Ukraine 5 AM Coalition devoted to documenting war crimes is launched in Ukraine," March 15, 2022, https://zmina.ua/en/event-en/ukraine-5-am-coalition-devoted-to-documenting-war-crimes-is-launched-in-ukraine/; Eliot Higgins, "These are the Cluster Munitions Documented by Ukrainian Civilians," Bellingcat, March 11, 2022, https://www.bellingcat.com/news/rest-of-world/2022/03/11/these-are-the-cluster-munitions-documented-by-ukrainian-civilians/.

discussions of the battalion that the company "dialed back" after receiving negative press.[23] Graphic violence policies also certainly apply, but appear to have been interpreted more loosely than in previous conflicts. However, the official Russian state media channels that have been shut down are also very important sources of evidence for thorough investigations. The only platform that has publicly committed to preserving content is Meta, and not in an announcement but rather as a comment to the BBC.[24]

Meta's willingness to consider the importance of its platform in OSI is encouraging, but unfortunately Meta is not the primary source of evidence in this conflict. Content from Syria, Sudan, Yemen, and other conflicts has been on Facebook and Instagram, but also Twitter and YouTube. Yet in this conflict, Tik Tok has emerged as a very important source. Another significant development is the importance of Telegram. This is new and poses challenges for OSI for a variety of reasons (including TVEC), as channels on the platform are sometimes spaces for far-right organizing. The German government threatened to ban Telegram but finally met with Telegram representatives early this year.[25] Since then, Telegram has shut down at least 64 channels in Germany.[26] Investigators focused on Ukraine report entire channels suddenly gone with no warning, which could be related to government pressure. Telegram has been very resistant in engaging with civil society organizations (much less governments). As a platform-focused body, GIFCT could encourage the company to engage more.

The situation in Ukraine has prompted more interest in OSI and makes it much more likely that the policy solutions proposed in this paper, in particular legislative solutions, will be adopted.

International Criminal Court

The ICC has been innovating around the use of OSI for several years now. In 2017, the ICC referred to Facebook videos in an arrest warrant for Mahmoud Mustafa Busayf Al-Werfalli.[27] This reference was groundbreaking for openly referring to social media content, but since then the ICC has continued to develop its approach to OSIs. In the case of The Prosecutor v. Ahmad Al Faqi Al Mahdi, the Court was presented "with a significant quantity of open-source evidence," including YouTube content.[28] In The Prosecutor v. Jean-Pierre Bemba Gombo, Aimé Kilolo Musamba, Jean-Jacques Mangenda Kabongo, Fidèle Babala Wandu, and Narcisse Arido, "the Prosecution argued that the defendant had bribed a witness to change their testimony. In support of this allegation, the Prosecution submitted evidence of a wire transfer as well as pictures from Facebook showing the two witnesses alleged to have been bribed together."[29]

23 Sam Biddle, "Facebook Allows Praise of Ukraine's Neo-Nazi Azov Battalion if It Fights Russian Invasion," The Intercept, February 24, 2022, https://theintercept.com/2022/02/24/ukraine-facebook-azov-battalion-russia/

24 James Clayton, "Are tech companies removing evidence of war crimes?" BBC, March 31, 2022, https://www.bbc.com/news/technology-60911099.

25 Reuters Staff, " Germany holds 'constructive' talks with Telegram, plans more," Reuters, February 4, 2022, https://www.reuters.com/article/germany-telegram-idUKKBN2K90Z9.

26 Deutsche Welle, " Telegram blocks over 60 channels in Germany – report," February 12, 2022, https://p.dw.com/p/46uZT.

27 Alexa Koenig, "Harnessing Social Media as Evidence of Grave International Crimes," Human Rights Center, Berkeley School of Law, October 23, 2017, https://medium.com/humanrightscenter/harnessing-social-media-as-evidence-of-grave-international-crimes-d7f3e86240d.

28 Róisín A. Costello, "International criminal law and the role of non-state actors in preserving open source evidence," Cambridge Journal of International Law, December 1, 2018, 268–283.

29 Id. (citing The Prosecutor v Jean-Pierre Bemba Gombo, Aimé Kilolo Musamba, Jean-Jacques Mangenda Kabongo, Fidèle Babala Wandu and Narcisse Arido (Judgment) ICC-01/05-01/13 (19 October 2016).

The ICC relies on OSIs because mutual legal assistance treaties (MLATs) and other formal tools to request evidence are largely not available to the Court. MLATs are agreements between two or more countries that "enable law enforcement authorities and prosecutors to obtain evidence, information, and testimony abroad in a form admissible in the courts of the Requesting State."[30] The ICC's rules of evidence are not as specific as (for example) the United States Federal Rules of Evidence, although these rules do help guide admission of evidence.[31] However, U.S. evidence rules would only allow the use of content obtained through OSI in very limited circumstances. In the ICC, evidence can be established through testimony in a more flexible way. The more markers of authenticity content has, the easier it is to use. This means that any level of preservation of and access to TVEC data could be helpful in prosecutions.

The SCA and GDPR

There are a variety of laws that apply to the retention and access to data. Unsurprisingly, due to platforms' geographical location, the most impactful of these laws are the United States Stored Communications Act (SCA), followed by the European General Data Protection Regulation (GDPR). The SCA is a federal law enacted as part of the Electronic Communications Privacy Act (ECPA) of 1986, an update on the Federal Wiretap Act of 1968. The SCA bans the disclosure of user data to third parties except under certain conditions. The SCA was updated by the 2018 CLOUD Act.

For purposes of this paper, the relevant provision of the SCA is Section 2702, which allows platforms to disclose customer communications to a law enforcement agency "if the contents... appear to pertain to the commission of a crime" or "to a foreign government pursuant to an order from a foreign government that is subject to an executive agreement that the Attorney General has determined and certified to Congress satisfied section 2523."

The SCA regulates just two categories of internet service provider (ISP), reflective of the state of technology at the time the statute was passed: "electronic communication service" (ECS) or a "remote computing service" (RCS). ECS is defined as "any service which provides to users thereof the ability to send or receive wire or electronic communications," and RCS is defined by the statute as "the provision to the public of computer storage or processing services by means of an electronic communications system." However, these distinctions come into play in Section 2703 concerning "required disclosure of customer communications or records." In order to require an ECS provider to disclose content that is in temporary electronic storage for 180 days or less, the government needs a search warrant. For an ECS provider to disclose content in storage for more than 180 days, or to make this request of an RCS provider, the government has three options: a search warrant, subpoena, or court order.

It should be noted that the applicability of any of these provisions to publicly posted content is questionable. Therefore, the SCA largely seems to play a role in the question of preservation and access due to platforms' concerns about potential liability. **That being said, once a platform deletes user-generated content, whether it is posted publicly or privately, people no longer have the ability to change the privacy settings. Thus,**

30 United States Department of Justice, "Mutual Legal Assistance Treaties of the United States" US DOJ, Criminal Division, Office of International Affairs, April, 2022, https://www.justice.gov/criminal-oia/file/1498806/download.

31 Alexa Koenig and Nikita Mehandru, "Open Source Evidence and the Criminal Court," Harvard Human Rights Journal, April, 2019, https://harvardhrj.com/2019/04/open-source-evidence-and-the-international-criminal-court/; International Criminal Court, "Rules of Procedure and Evidence," 2013, https://www.icc-cpi.int/sites/default/files/RulesProcedureEvidenceEng.pdf.

regardless of the applicability of the GDPR platforms should treat even public posts with care.

The European Union (EU) GDPR came into effect in 2018. The GDPR regulates the collection and processing of personal data by organizations and companies like technology companies as well as government agencies. It requires that companies that process data ensure there are specific purposes for doing so and make this clear to individuals when collecting their data (known as purpose limitation). Companies must also establish time limits to erase or review the data stored (storage limitation) and respond to requests from individuals exercising their data protection rights free of charge within 1 month of receipt. Under the GDPR, data subjects have the following privacy rights: to be informed; have access; seek rectification; restrict processing; ensure erasure and data portability; object; and exercise their rights in relation to automated decision-making and profiling. To be protected under the GDPR, a data subject has to be either a citizen of the EU or simply located in the EU, regardless of national identity.

The GDPR could apply to the preservation and access of relevant data. Currently, platforms' terms of service refer only to law enforcement investigations and not the other potential uses of data considered by this paper. It should be noted that personal or household activities, national security, and law enforcement are exempt from the EU GDPR. As noted, the Swedish and German courts that used open-source data to prosecute war crimes did not run into GDPR issues. However, more analysis is needed to determine whether and how the GDPR might apply to data shared with the ICC and United Nations.

High-level principles for solutions

Discussions about the need to solve the problem of disappearing human rights content have been going on for many years now, but thanks to the visibility of social media's role in Myanmar and the increasingly obvious utility of this content, policymakers are taking this issue seriously.

Many groups have been working on this topic. One is an informal coalition composed of Human Rights Watch, WITNESS, Mnemonic, and the Berkeley Human Rights Center. Starting in January 2020, this group hosted a series of workshops to understand the views of various stakeholders. These workshops included a wide range of participants, including lawyers, content moderation experts, representatives of community archives from conflict zones, and privacy experts. The goal was not to propose one specific solution, but rather to consider the principles solutions ought to follow.

In the end, the informal coalition gathered at these workshops reached a consensus position: while there is no one size fits all solution, there are a few things that should be considered in any solution. The most pressing principle to consider is the tension between privacy/security and the purpose of preserving and providing access to data. This is reflected in the principle of necessity and proportionality in international human rights law.

There was no consensus among workshop attendees about how to articulate principles for solutions. However, many of the resources referenced in these workshops inform this paper. In addition to concerns about privacy and security, there were a few other overarching concerns, spanning responsible retention and stewardship of data (including transparency), holding data for limited and clearly defined purposes, and having methods to ensure remedies for parties negatively impacted by solutions. One seldom-noted principle that should be recognized and implemented is the idea that impacted communities should be centered in discussions about

possible solutions.[32]

As noted above, the tension between privacy and security and access to justice is particularly important to consider. While it may be easy to simply ask platforms to preserve data, there are always potential harms in preserving and providing access to PII (regardless of whether it was originally posted in a public or private forum).

PII is data that could be used to identify an individual. The GDPR provides a very helpful definition: "personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person."[33]

The vast majority of people whose PII is included in content are not only not accused of any crime but are uniquely vulnerable. Their privacy, security, and even safety can be impacted by the retention of data. Solutions should weigh these potential harms against how the data will be used and who it benefits. PII that is stored by companies should be considered accessible by governments and law enforcement agencies, including those government actors that might misuse the data, or use it in service of human rights violations. This is not hypothetical when it comes to social media data. For example, in 2017 the Egyptian government ramped up its targeting of LGBTQ people after images of attendees waving a rainbow flag at a Mashrou Leila circulated on social media. The Egyptian government has a unit dedicated to arresting and prosecuting LGBTQ people that uses social media content as evidence.[34]

These concerns are in no way limited to authoritarian governments. The United States and the EU have also unfairly targeted vulnerable communities using social media posts and associated data. For example, under President Donald Trump, non-immigrant visa applications were updated to ask applicants for their social media handles, in service of Trump's colloquially named "Muslim Ban."[35] In a case challenging this requirement, plaintiffs said "The Registration Requirement is the cornerstone of a far-reaching digital surveillance regime that enables the U.S. government to monitor visa applicants' constitutionally protected speech and associations not just at the time they apply for visas, but even after they enter the United States."[36] It was expected that President Joe Biden would do away with this practice, but instead his administration has recommended expanding it.[37] In addition

32 See, eg, Carroll, S.R., Garba, I., Figueroa-Rodríguez, O.L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J.D., Anderson, J. and Hudson, M., "The CARE Principles for Indigenous Data Governance." Data Science Journal, November 4, 2020, 19(1), p.43. DOI: http://doi.org/10.5334/dsj-2020-043; Society of American Archivists, "SAA Core Values Statement and Code of Ethics." Approved by the Society of American Archivists, last revised August 2020, https://www2.archivists.org/statements/saa-core-values-statement-and-code-of-ethics.

33 Article(1), REGULATION (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

34 Declan Walsh, "Egyptian Concertgoers Wave a Flag, and Land in Jail," New York Times, September 27, 2017, https://www.nytimes.com/2017/09/26/world/middleeast/egypt-mashrou-leila-gays-concert.html.

35 United States Department of State, "Frequently Asked Questions on Social Media Identifiers in the DS-160 and DS-260," June 4, 2019, https://travel.state.gov/content/dam/visas/Enhanced%20Vetting/CA%20-%20FAQs%20on%20Social%20Media%20Collection%20-%206-4-2019%20(v.2).pdf.

36 Doc Society et al v. Pompeo, COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF, filed December 12, 2019, https://www.brennancenter.org/sites/default/files/2019-12/Complaint%20Doc%20Society%20v%20Pompeo.pdf.

37 Anna Diakun and Carrie DeCell, "Why is the U.S. still probing foreign visitors' social media accounts?," Washington Post, April 26, 2022, https://www.washingtonpost.com/outlook/2022/04/26/social-media-surveillance-us-visas-state/.

to the U.S. government's use of social media, myriad federal agencies use social media monitoring in ways that have led to false arrests and other human rights violations.[38] Providing these agencies with a vast store of data should not be the end result of attempts to preserve human rights documentation and other important data.

European law enforcement agency handling of data raises similar concerns. In January of this year, the European Data Protection Supervisor (EDPS) ordered Europol "to delete data concerning individuals with no established link to a criminal activity (Data Subject Categorisation)."[39] Sensitive data was "sampled from asylum seekers never involved in any crime."[40]

A note on data related to attacks with an online component

The livestreamed March 2019 attack on mosques in Christchurch, New Zealand and the livestreamed May 2022 attack on a grocery store in a predominantly Black neighborhood of Buffalo, New York exemplify the kinds of incidents that GIFCT and the Christchurch Call to Eradicate Terrorist and Violent Extremist Content Online are meant to respond to. GIFCT and New Zealand government officials have created the Content Incident Protocol (CIP) and the Crisis Response Protocol (respectively) for incidents such as these. Both of these protocols are meant to allow situational information sharing, and the CIP allows companies to share hashes of content created by perpetrators or their accomplices (the CIP in particular is limited in scope to this perpetrator content). What the CIP does not necessarily do is ensure that governments have access to technical information about this content that could help them understand the online component of these attacks.

After the Christchurch shooting, the New Zealand government wanted access to information such as how many times the video had been viewed and the location of users who viewed it. They were not able to obtain that information. It is essential that governments not bypass due process requirements for access to data and further research is needed to address existing relevant legal frameworks for sharing such information. It should be noted that even non-PII could help aid in such post-incident investigations. There seem to be technical limitations to how much information companies can provide, as well as confusion within companies about who should be providing this information. The CIP, as well as the related Christchurch Call Crisis Response Protocol, helps to define points of contact. However, this information is not sufficient for large companies like Meta and Google – these companies need to determine how to provide non-PII and who should be responsible for the necessary steps. GIFCT should aid in this by ensuring that due process requirements and applicable legal frameworks are clarified and made available to small companies.

Company efforts: Twitter's research consortium

It is worth noting that some platforms are already providing more data to researchers than others. Twitter stands out as one of the platforms with the most accessible data. The company has always been notable in

38 Rachel Levinson-Waldman, Harsha Panduranga and Faiza Patel, "Social Media Surveillance by the U.S. Government," Brennan Center for Justice, February 7, 2022, https://www.brennancenter.org/our-work/research-reports/social-media-surveillance-us-government.

39 European Data Protection Supervisor, "EDPS orders Europol to erase data concerning individuals with no established link to a criminal activity," January 10, 2022, https://edps.europa.eu/press-publications/press-news/press-releases/2022/edps-orders-europol-erase-data-concerning_en.

40 Apostoli Fotiadis et al., "'A data 'black hole': Europol ordered to delete vast store of personal data," The Guardian, January 10, 2022, https://www.theguardian.com/world/2022/jan/10/a-data-black-hole-europol-ordered-to-delete-vast-store-of-personal-data.

providing meaningful API access, but it also has an academic researcher program that allows greater access. Furthermore, in December 2021, the company announced that it would be launching the Twitter Moderation Research Consortium to "provide comprehensive data about attributed platform manipulation campaigns to members of the consortium, who may independently choose to publish their findings on the basis of the data we share and their own research."[41] The consortium has just launched and is available to a limited number of academic researchers. **This** effectively excludes archives like Mnemonic, **and hopefully Twitter will determine standards that do allow broader access**. However, as it grows, Twitter's work in defining standards for access and protection of users could help to inform further GIFCT research on this topic.

What we know about how companies handle data

In the course of conducting research for this paper, a survey was sent to social media companies asking them to explain how they determined where GDPR and SCA applied to the data they hold, what their data retention and deletion practices are, and what their law enforcement request procedures are. Unfortunately, only one company provided a response, leaving us to guess how exactly GIFCT member companies are handling data.

It is impossible to get a clear picture of how all GIFCT member companies handle data because most of them do not provide specific time frames for how long it takes them to delete data-including backup copies after a user deletes it, nor how companies determine what laws apply to specific users (in particular the GDPR).

However, we were able to find some information from public posts and policies about how member companies handle data. Currently, it appears that they have few limitations on storing data. Meta says, "We store data until it is no longer necessary to provide our services and Meta Products, or until your account is deleted – whichever comes first.... When you delete your account, we delete things you have posted, such as your photos and status updates, and you won't be able to recover that information later."[42] It does not provide any specific time frame. Twitter's policy states, "We keep your profile information and content for the duration of your account. We generally keep other personally identifiable data we collect when you use our products and services for a maximum of 18 months." Twitter's policy explains that a user's account information will be held for up to 30 days, and that "[w]here you violate our Rules and your account is suspended, we may keep the identifiers you used to create the account (i.e., email address or phone number) indefinitely to prevent repeat policy offenders from creating new accounts."[43] Google's privacy policy says that they delete content when it is deleted by users, but that the whole process "generally takes around 2 months from the time of deletion," and that data on encrypted backup servers "can remain on these systems for up to 6 months."[44] Links to other GIFCT member privacy policies can be found in the footnotes, but the pattern is clear: these policies do not address how platforms manage data they took down themselves. While it is clear that data is not deleted instantly, the policies leave significant room for platforms to selectively preserve content where appropriate – **keeping in mind that once a platform deletes content, users no longer have the option to change privacy settings.**

41 Gadde Vijaya and Yoel Roth, "Expanding access beyond information operations," Twitter Blog, June 7, 2022, https://blog.twitter.com/en_us/topics/company/2021/-expanding-access-beyond-information-operations-.

42 Meta, "Data Policy," January 4, 2022, https://www.facebook.com/about/privacy/update. Note that the policy is labelled differently for the Meta Platforms Ireland Limited, which processes data for European Union residents, and Meta Platforms Inc. This information is under "Data retention, account deactivation and deletion" for Meta Ireland and "How can I manage or delete information about me?" for Meta.

43 Twitter, "Privacy Policy," June 10, 2022, https://twitter.com/en/privacy.

44 Google, "How Google retains the data we collect," June 10, 2022, https://policies.google.com/technologies/retention.

No platform has publicly committed to retaining specific types of data after the activation of a CIP, nor in times of crisis where it is widely known that human rights documentation is amassing online (e.g., in Ukraine, but also Iran, Palestine, Sudan, and other crises). **Considering the legal, privacy and security implications raised by storing user data this is understandable, but it's clear that a solution is needed as soon as possible.**

Recommendations

Interviews and literature review during the course of this research reinforced the conclusion that it would be a mistake at this time to simply ask platforms to retain all data or hand it over to most non-governmental parties. Instead, there are a number of limited steps that should be taken, and further research undertaken.

First, there is one problem that has a clear solution: the need for a mechanism to allow the ICC and UN investigative bodies to request data. The most likely form this mechanism would come in is a very limited exception to the SCA disclosure limitations. This would likely have to be a direct amendment to the SCA, and it should be written in the narrowest possible way. It should provide an exception only for content relevant to a limited set of international human rights law violations, including war crimes, crimes against humanity, and genocide. Scholars at Yale and Boston College have proposed just such a solution.[45] However, due to the massive privacy implications of tampering with the SCA, this paper suggests that any amendment should be written in the most narrow way possible, and should comport with due process requirements. It could require the ICC and UN to apply with a magistrate judge in line with Rule 41 of Federal Rules of Criminal Procedure.[46] It could establish standards for a data request that conform as much as possible to 4th Amendment warrant requirements, which require applications for warrants to be justified by probable cause, supported by oath or affirmation, and in particular describe the place to be searched and the persons or things to be seized. Additionally, where necessary, platforms should refine their policies to clarify that data could be used for the purpose of prosecuting the same limited set of international crimes in order to comply with the GDPR. Further steps may be needed to comply with the GDPR, and GDPR experts should be consulted.

Further research is needed to determine exactly what gaps exist for governments to request data from platforms about content after a terrorist attack with an online component. The purpose of the GIFCT CIP and the Christchurch Crisis Response Protocol is partly to enable information sharing, but it should not be done in a way that bypasses due process. Some of the metadata governments are interested in is not necessarily PII: for example, platforms could provide governments with information about the locations of accounts or the number of times a piece of content was viewed without implicating privacy concerns. GIFCT is well-positioned to address this topic, and it should be the focus of the working group next year.

Similarly, further investigation and analysis is needed to determine how to provide increased civil society access to data, either by directly providing copies of user-generated content depicting human rights violations or providing data for other research purposes (such as understanding the spread of misinformation). Twitter's experience could be very valuable here.

45 Joshua Lam and David Simon, "To Support Accountability for Atrocities, Fix U.S. Law on the Sharing of Digital Evidence," Just Security, April 20, 2022, https://www.justsecurity.org/81182/to-support-accountability-for-atrocities-fix-u-s-law-on-the-sharing-of-digitial-evidence/; Rebecca J. Hamilton, "Platform-Enabled Crimes," Boston College Law Review, November 12, 2021, https://ssrn.com/abstract=3905351 or http://dx.doi.org/10.2139/ssrn.3905351.

46 FRCP 41 "Search and Seizure," https://www.law.cornell.edu/rules/frcrmp.

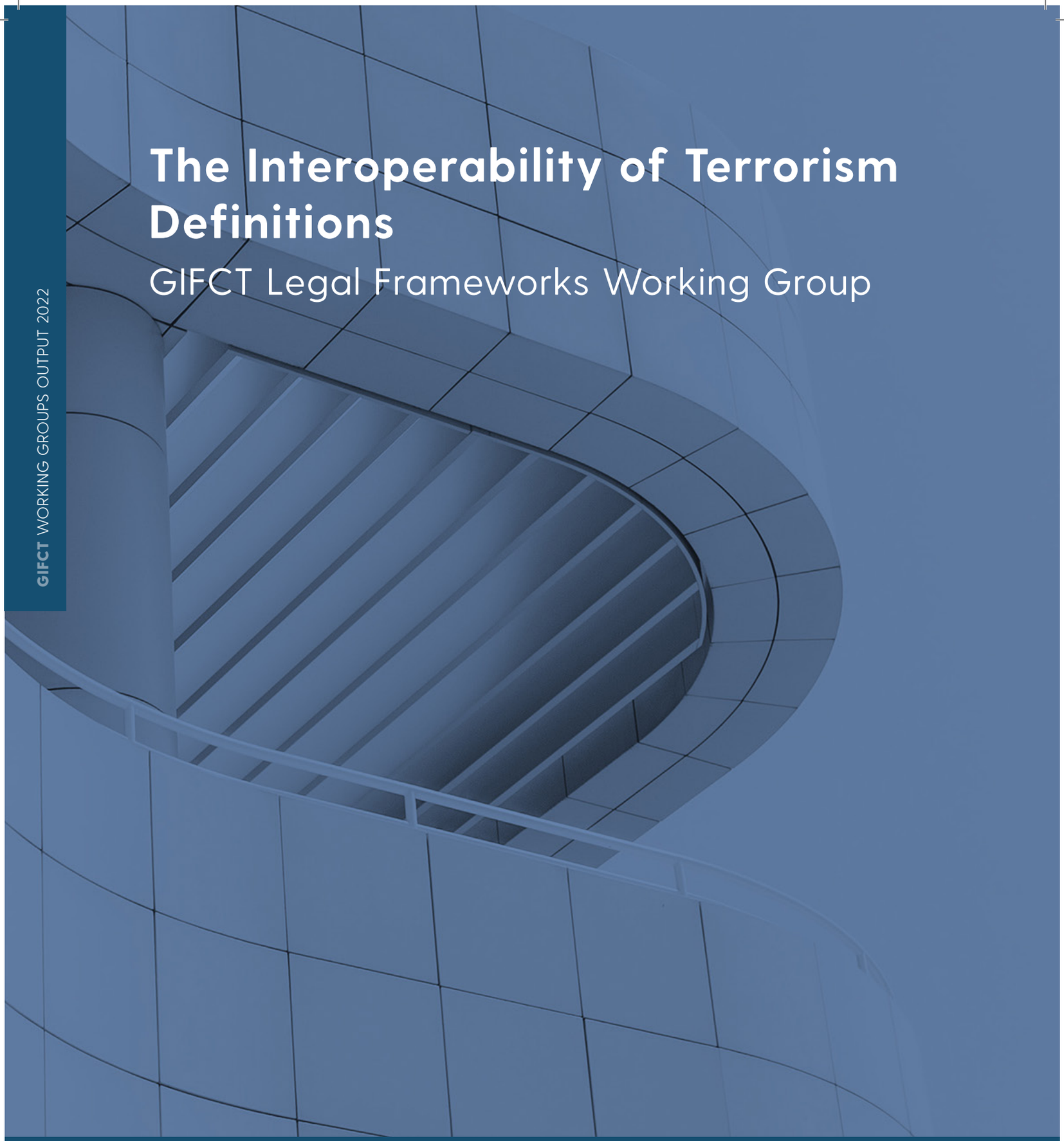
Finally, this research reinforced the need for increased transparency from platforms in their privacy policies or terms and conditions. Platforms should clarify how long to retain data when a user has deleted it, and explain exactly how long they handle data from content that they themselves have removed. They should also explain with more clarity what legal frameworks they apply to what users. Broadly, they should also commit (or recommit) to the Santa Clara Principles On Transparency and Accountability in Content Moderation, **which provides in-depth and operationalizable standards for transparency from platforms**.[47]

Ultimately, it is clear that OSIs are a compelling necessity that should be addressed as soon as possible. Similarly, in the wake of the Buffalo shooting, it is clear that more information about how perpetrator content travels in the media ecosystem is necessary to understand the online aspect of such violent attacks. These are achievable goals that should be prioritized by civil society organizations, governments, and companies.

Interview list

Hadi al Khatib (Syrian Archive/Mnemonic), David Shanks (former), Lindsay Freeman (UC Berkeley Human Rights Center), Nick Waters (Bellingcat) , Aaron Zelin (Jihadology/Brandeis), Sun Kim (IIMM but interviewing in her personal capacity), Yvonne McDermot (Swansea), Nathaniel Raymond (Yale), Libby McAvoy (Mnemonic), Anonymous activists

47 The Santa Clara Principles on Transparency and Accountability in Content Moderation, 2021, https://santaclaraprinciples.org/.



GIFCT WORKING GROUPS OUTPUT 2022

The Interoperability of Terrorism Definitions

GIFCT Legal Frameworks Working Group

GIFCT
Global Internet Forum to Counter Terrorism

Dr. Katy Vaughn
Swansea University

Introduction

It is a condition of GIFCT membership that companies must have policies that "explicitly prohibit terrorist and/or violent extremist activity."[1] However, where policies exist, many tech platforms, companies, and other actors lack a coherent and consistent approach to defining terrorist and violent extremist content (TVEC) that aligns with relevant sources of law and human rights standards. Among the founding members of the GIFCT, only one platform (Meta) has a definition of terrorism, with the others relying on designation lists and definitions of violent extremism.

Approaches to defining TVEC underpinned by international and national designation lists will cover material from specific identified groups and individuals and requires frequent updates. More general and behavioral approaches to defining TVEC may lack consistency in scope and application and are generally harder to monitor and apply. Under algorithm-based models, there would be flagged material that meets the definitions applicable everywhere, material that meets most of the definitions in multiple regimes, material that meets the definitions in some regimes but not most, material that does meet the definitions in all but a few outlier regimes, and material that would not reasonably meet the definitions in any regime.

This may impede the proper application of content management policies about suspected TVEC on the platform and interfere with companies' capacity to operate effectively and efficiently with external actors. While GIFCT creating a shared definition of terrorism is met with skepticism—as this task more appropriately sits with governments and international consensus-building forums—the recent GIFCT Human Rights Impact Assessment recognized the value in creating a common understanding of TVEC.[2] It is in the interest of GIFCT member companies to promote a move towards interoperability, with approaches to defining TVEC that are coherent and consistent throughout multiple regimes and jurisdictions. As identified by BSR, the benefits could include "pushing back against overbroad definitions" which could, in turn, establish a "bulwark" against definitions that present risks to the freedom of expression, "improving the capability of smaller companies," and "improving shared awareness of the relationship between human rights and terrorist and violent extremist content."[3]

The GIFCT Legal Frameworks Working Group (LFWG) is seeking to facilitate an expansion of the understanding of the possible implications for companies of the current level of (in)coherence regarding TVEC across commercial, national, regional, and international definitions of terrorism. Developing policy guidance that would facilitate broadly compatible approaches towards a definition of TVEC—in line with a range of potentially applicable legal regimes and jurisdictions—first requires an understanding of how interoperable and coherent existing definitions are.

The premise of this approach is that there are standard minimums, thresholds, or tests to establish whether content meets the standard of TVEC. Where there is incoherence that impacts significantly on human rights, coherent norms need to be developed.

1 GIFCT, "Membership," n.d., https://gifct.org/membership/.

2 BSR, "Human Rights Impact Assessment: Global Internet Forum to Counter Terrorism," 2021, https://www.bsr.org/en/our-insights/report-view/human-rights-impact-assessment-global-internet-forum-to-counter-terrorism.

3 BSR, Human Rights Impact Assessment.

Objectives

This paper aims to compare a broad sample of definitions of terrorism and violent extremism from tech companies, national laws, regional legislation, international legal instruments, and international human rights law standards. The objective of this comparison is to identify the level of interoperability and (in)coherence in and between the definitions examined. This analysis will underpin future efforts to reduce inconsistent and conflicting definitions, recommend suitable minimum standards or principles underpinning definitions of TVEC, and improve cross-sectoral coherence.

Method

Definitions were collected from international instruments (four), regional intergovernmental instruments (five), domestic statutory definitions (four jurisdictions), the publicly available policies of the GIFCT founding members (four), and the model definition of terrorism put forward by the Special Rapporteur on counterterrorism and human rights. Academic definitions were also considered, including Schmid and Jongman's, which is based on an academic consensus resulting from a study conducted in 1989.[4] Studies since have utilized Schmid and Longman's methodology examining selected legal definitions against the word categories triggered by the academic consensus.[5] Instead of examining against Schmid and Jongman's word categories, this study examines a broader set of definitions of terrorism and identified categories based on an initial analysis of the selected definitions and the standard features found in these definitions. This was used as the basis of a comparative analysis to identify existing interoperability between them.

The definitions were selected by members of GIFCT's LFWG, drawing from existing geographical areas of expertise. The U.S. definitions were important given GIFCT's founding members are based in the U.S. Australia's definition is indicative of other Commonwealth country approaches such as the U.K., New Zealand, and Canada. France provides an example from a different type of legal system, as does Indonesia.

Some national definitions and the platform definitions for violent extremism were considered where available. It was noted that violent extremism was not legally defined. Some sources of law for violent extremism exist but are not referred to as violent extremism legislation. Some jurisdictions covered in this paper do have non-terrorism legislation that would capture violent extremist material, such as incitement to violence laws (Australia),[6] laws regarding the recording and dissemination of violent images (France),[7] and online safety laws targeting the publication of abhorrent violent material (Australia).[8] Therefore, terrorism definitions were the priority of this first phase of analysis. However, violent extremism seems to have emerged as an important category for platforms in response to the difficulties and complexities of navigating terrorism law.

4 A. P. Schmid and A. J. Jongman, Political Terrorism: A New Guide to Actors, Authors, Concepts, Data Bases, Theories, & Literature (Transaction Books, 2005): 5–6.

5 Jessie Blackbourn, Fergal F. Davis, and Natasha C. Taylor, "Academic Consensus and Legislative Definitions of Terrorism: Applying Schmid and Jongman," Statute Law Review 34, no. 3 (October 2013): 239–261, https://doi.org/10.1093/slr/hms041.

6 For example, in New South Wales Publicly threatening or inciting violence is an offense under section 93Z of the Crimes Act 1900.

7 Article 222-33-3 of the Code Pénal.

8 Section 474.31 of the Criminal Code 1995; Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019, the Criminal Code is mirrored by application provisions in the Online Safety Act 2021 (Cth). See also the Information and Electronic Communications Law in Indonesia (Undang-undang No 11 tahun 2008 tentang Informasi dan Transaksi Elektronik). In Indonesia, there is also a presidential decree (peraturan presiden) No 7 year 2021 on National action plan on prevention and countering violent extremism.

Consequently, the mapping of sources of law for violent extremism (beyond terrorism) would be an important further phase of analysis as part of GIFCT efforts to ensure that platform categories and definitions align with the law.

The research uses qualitative analysis (specifically content analysis) to identify themes, patterns, and relationships between the definitions. Primary data was examined and then compared to phenomena highlighted in relevant academic literature to identify coherence, interoperability, and impact.

The research was conducted over four stages:

1. Collate a broad sample of definitions of terrorism and violent extremism used by international bodies, regional organizations, nation-states, and companies.
2. Undertake a comparative analysis to identify similarities and differences between standard features across the definitions.
3. Analyse the degree of coherence between the definitions and against human rights standards such as the principle of legality, necessity, and proportionality. Identify the areas of greatest coherence and areas of divergence and incoherence.
4. Evaluate the most harmful areas of divergence and incoherence in relation to human rights and community impacts.

This paper begins by introducing the standard features and core requirements of definitions of terrorism and identifying and explaining the relevant human rights standards. It will then highlight the areas of coherence and divergence across the definitions (and the impact on rights protection) before providing an analysis of some of the most harmful areas of divergence.

Definitions of Terrorism and Violent Extremism

In the absence of an internationally agreed definition of terrorism, some companies and organizations have resorted to relying on terrorist designation lists as a basis for policies on responding to TVEC online. Initially, GIFCT's hash-sharing database was limited to material associated with designated organizations on the United Nations Security Council's Consolidated Sanctions List, with the view to focusing on a narrow set of content.[9] Of the definitions of companies included in this study, Twitter, YouTube, and Microsoft avoid including a definition of "terrorism" and instead state that their policies are informed by terrorism designation lists. Twitter and YouTube refer to national and international terrorist designation lists without referring to a specific list; Microsoft relies on the UN Consolidated Sanctions List.

Reliance on international and domestic terrorist designation lists to define TVEC and design policies to counter TVEC online presents the danger that such policies will reflect "broader discrimination and bias in the

9 GIFCT, "Broadening the GIFCT Hash-sharing Database Taxonomy: An Assessment and Recommended Next Steps," July, 2021, https://gifct.org/wp-content/uploads/2021/07/GIFCT-TaxonomyReport-2021.pdf.

counterterrorism field"[10]—specifically disproportionately focusing on self-declared Islamist terrorist organizations and not right-wing extremist groups. This in turn can have a disproportionate effect on Muslim and Arab communities.[11] There is also the difficulty in identifying right-wing terrorist groups or organizations, which can make utilizing designation difficult, as groups are "ideologically and organizationally fragmented."[12] In contrast to the "group structure" of organizations such as IS, the extreme right is "not dominated by one or even a small number of groups."[13] This is evident in recent far-right terrorist attacks carried out by individuals who are difficult to identify as "members or even supporters of formal groups."[14] GIFCT responded to these concerns and has identified that it could expand its taxonomy based on a "behavioral and content-focused approach" and organizations.[15]

This is not to say that in the absence of a universal definition of terrorism there are no advantages (particularly for companies) to list-based approaches. This underpins Tech Against Terrorism's development of the Terrorist Content Analytics Platform (TCAP), which alerts platforms to content associated with designated terrorist organizations and is stated to be an approach grounded in the rule of law.[16] TCAP does include content created by both designated Islamist terrorist organizations and designated far-right terrorist organizations.[17]

While using designated terrorist organizations as the basis for identifying content provides certainty for companies, there are inherent difficulties including a rule of law perspective on the processes used by states and intentional bodies to designate groups as terrorist organizations.[18] The lack of due process and transparency are well documented, particularly at the international level, including listing procedures carried out by the United Nations.[19] Therefore, it is concerning if companies are relying on list-based approaches as a basis for designing policies on moderating TVEC on their platforms and services. In addition, removing content solely based on it being associated with a designated organization is problematic. This is illustrated by a decision by Meta's Oversight Board, which overturned Meta's original decision to remove an Instagram post encouraging people to discuss human rights concerns relating to the solitary confinement of a founding member of the Kurdistan

10 GIFCT, Hash-sharing Database Taxonomy; BSR, Human Rights Impact Assessment.

11 Svea Windwehr and Jillian C. York, "One Database to Rule Them All: The Invisible Content Cartel that Undermines the Freedom of Expression Online," Electronic Frontier Foundation, August 27, 2020, https://www.eff.org/deeplinks/2020/08/one-database-rule-them-all-invisible-content-cartel-undermines-freedom-1.

12 Report of the Secretary General, "Activities of the United Nations system in implementing the United Nations Global Counter-Terrorism Strategy," A/76/729, January 29, 2021.

13 Maura Conway, "Routing the Extreme Right," The RUSI Journal 165, no. 1 (28 February 2020): 108–113, https://doi.org/10.1080/03071847.2020.1727157. For discussion on the lack of clarity in general on what constitutes a "terrorist group," see Brian J. Phillips, "What Is a Terrorist Group? Conceptual Issues and Empirical Implications," Terrorism and Political Violence 27, no. 2, (2015): 225–242, https://doi.org/10.1080/09546553.2013.800048.

14 Conway, "Routing the Extreme Right."

15 GIFCT, Hash-sharing Database Taxonomy.

16 Tech Against Terrorism, "The Terrorist Content Analytics Platform and Transparency By Design," VOX-Pol (Blog), November 11, 2020, https://www.voxpol.eu/the-terrorist-content-analytics-platform-and-transparency-by-design/.

17 Terrorist Content Analytics Platform, "Inclusion Policy," 2021, https://www.terrorismanalytics.org/policies/inclusion-policy.

18 Report of the Eminent Jurists Panel on terrorism, counter-terrorism and human rights, International Commission of Jurists, 2009.

19 Report of the Eminent Jurists; Report of the United Nations High Commissioner for Human Rights on the protection of human rights and fundamental freedoms while countering terrorism, A/HRC/16/50, December 15, 2010; James Cockayne and Rebecca Brubaker, "Due Process in the UN Targeted Sanctions: Old Challenges, New Approaches," United Nations University Conference Proceedings, March, 2020, https://collections.unu.edu/eserv/UNU:7615/DueProcess_ConferenceBrief.pdf; Gavin Sullivan, The Law of the List: UN Counterterrorism Sanctions and the Politics of Global Security Law (Cambridge University Press, 2020).

Workers' Party (PKK)—a designated terrorist organization.[20]

Given the inherent difficulties with the list-based approach, general definitions of terrorism have emerged. This is the approach of most of the definitions included in this study (other than those employed by Twitter, YouTube, and Microsoft), and will form the basis of the analysis of coherence between them based on standard features/core requirements. These include an act of violence, causing intentional harm, with the intention to impact a target beyond the immediate victims (wider audience), with underlying motives, and including express exemptions.

At the international level, there has been little attempt to define violent extremism. Given the absence of a universally agreed definition, vague and broad definitions have emerged at the national level. Interestingly, while Twitter and YouTube have avoided defining terrorism, they do include definitions of or references to violent extremism. These will be examined alongside the three found at the regional and national level (Shanghai Convention, U.S., and Australia) as part of this study. The development of policies that have the potential to negatively impact people's lives based on a term that has little legislative basis can be even more dangerous for human rights than the term 'terrorism.'[21]

Terrorist and Violent Extremist Content

This paper is focused on an examination of definitions of terrorism or terrorist acts. This is relevant to tech companies defining and designing policies in response to TVEC. However, this paper provides only a starting point. The definition of terrorist content should not depart from the definition of a terrorist act, and it is argued here that there is an advantage in greater interoperability in this regard. However, the difference between a terrorist act and terrorist content relates to the nexus among the content, the terrorist act, and the standard of proof that may be required.

In terms of the intersection between the terrorist act and terrorist content, a tech company policy may extend to content that shows, incites, glorifies, or instructs a terrorist act. This paper contends with the key elements of the definition of a terrorist act but does not assess the proof required to establish showing, inciting, glorifying, or instructing a terrorist act. Therefore, the intent component of the terrorist act will be relevant to determining if an event is a terrorist act, but not determinative or material towards the intent of the poster of content (unless the poster is engaged in the terrorist act themselves). Defining terrorist content and these associated questions are identified as a further area of research.

Human Rights Standards

The difficulties in defining terrorism and the implications of imprecise and overbroad definitions are well-known. They carry the risk of potential deliberate misuse, being unintentionally misapplied to acts that are not

20 "Oversight Board overturns original Facebook decision: Case 2021-006-IGUA," https://www.oversightboard.com/news/187621913321284-oversight-board-overturns-original-Meta-decision-case-2021-006-ig-ua/. Facebook had misplaced policy guidance including this exemption.

21 Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, A/HRC/31/65, April 29, 2016.

normally considered terrorism, and unintended consequences such as human rights violations.[22] The lack of an international consensus on a definition of terrorism coupled with the highly politicized context exacerbates these difficulties.[23] However, in the absence of an internationally agreed definition, human rights and the rule of law standards can counter these negative impacts,[24] and provide a framework in which to hold both states and companies accountable across national borders.[25]

The Principle of Legality

The principle of legality requires that measures taken by states must be lawful and that where a measure restricts human rights, those restrictions must be defined clearly and precisely to enable individuals to predict what limits the measure places on their behavior—and to ensure the measures taken do not encompass conduct which allows the law to operate beyond its intended purposes and scope.

Article 15, paragraph 1 of the International Covenant on Civil and Political Rights (ICCPR) provides that,

> **No one shall be held guilty of any criminal offence on account of any act or omission which did not constitute a criminal offence under national or international law at the time it was committed. Nor shall a heavier penalty be imposed than the one that was applicable at the time when the criminal offence was committed. If, subsequent to the commission of the offence, provision is made by law for the imposition of the lighter penalty, the offender shall benefit thereby.**

Article 15 is a non-derogable right.[26] Similar provisions are included in the European Convention on Human Rights (ECHR)[27] and the American Convention on Human Rights (ACHR).[28] These provisions embody the principle of legality, which requires states to give reasonable notice of any conduct that will attract criminal punishment. In compliance with Article 15 ICCPR, any prohibition on terrorist conduct must be prescribed by law, so the prohibition must be framed in a law that is "adequately accessible so that an individual has a proper indication of how the law limits his or her conduct; and the law is formulated with sufficient precision so that the individual can regulate his or her conduct."[29]

In the context of countering TVEC online, measures taken risk interference with the right to freedom of expression. Article 19 ICCPR guarantees the right to freedom of expression, which includes the right to receive information and the right to hold opinions without interference. The principle of legality is expressly included in the provision, stating that Article 19 rights can only be subject to restrictions where these are provided by law.

22 Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, "Ten areas of best practice in countering terrorism," A/HRC/16/51, December 22, 2010.

23 BSR, Human Rights Impact Assessment.

24 Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, E/CN.4/2006/98, December 28, 2005.

25 Report of the Special Rapporteur on the promotion and protection of the right to freedom of expression, A/HRC/38/35, April 6, 2018.

26 Article 4(2) ICCPR.

27 Article 7 ECHR.

28 Article 9 ACHR.

29 Report of the Special Rapporteur, E/CN.4/2006/98.

Similar provisions are included in Article 10 ECHR and Article 13 ACHR, which also expressly include the principle of legality. Restrictions on the right must, "be adopted by legal processes and limit government discretion in a manner that distinguishes between lawful and unlawful expression with 'sufficient precision.'"[30] This is also the case in relation to other relevant rights when considering measures regulating content online such as the right to freedom of thought, conscience, and religion,[31] the right of assembly,[32] and the right to privacy.[33]

It is useful here to refer to the work of Hardy and Williams, which established three criteria for examining whether domestic statutory definitions of terrorism are appropriate.[34] The first two criteria are encompassed by the principle of legality. In this sense, they advanced a legal definition of terrorism:

1. "... should be sufficiently clear and precise to give reasonable notice of the kinds of conduct it prohibits";[35] and
2. "... should not encompass conduct which allows legislation to operate outside its intended purposes."[36]

Therefore, it is important to assess definitions of terrorism and violent extremism against these standards. The principle of legality requires that the definitions be clear, precise, and narrowly focused. In the counterterrorism landscape, definitions must define terrorism and violent extremism such that counterterrorism measures are restricted to countering conduct that is truly terrorist in nature. In the context of responses to TVEC online, content moderation rules should be clear and specific to "enable users to predict with reasonable certainty what content places them on the wrong side of the line."[37]

Necessity and Proportionality

Where definitions of terrorism do not comply with the principle of legality and consequently fail to restrict counterterrorism measures to conduct which is terrorist in nature, there is also the danger that such measures will not comply with the principles of necessity and proportionality.[38] These are important principles placing limits on unjustified interferences with fundamental human rights and freedoms. The right to freedom of expression,[39] the right to freedom of thought, conscience, and religion,[40] and the right of assembly[41] are all qualified rights. However, any restrictions, in addition to needing to comply with the principle of legality, must be necessary and

30 Report of the Special Rapporteur, A/HRC/38/35, 6 April 2018.

31 Article 18 ICCPR, Article 9 ECHR, Article 12 ACHR.

32 Article 21 ICCPR, Article 11 ECHR, Article 15 ACHR.

33 Article 17 ICCPR, Article 8 ECHR, Article 11 ACHR.

34 Kieran Hardy and George Williams, "What is 'Terrorism'? Assessing Domestic Legal Definitions," UCLA Journal of International Law and Foreign Affairs 16 (2011): 77–162.

35 Kieran Hardy and George Williams, "What is 'Terrorism'? Assessing Domestic Legal Definitions," p.81.

36 Kieran Hardy and George Williams, "What is 'Terrorism'? Assessing Domestic Legal Definitions," p.81.

37 Report of the Special Rapporteur, A/HRC/38/35, para 46.

38 Report of the Special Rapporteur, A/HRC/16/51.

39 Article 19 ICCPR.

40 Article 18 ICCPR.

41 Article 21 ICCPR

proportionate to the legitimate aim. Therefore, the legal response to terrorism and violent extremism must not be disproportionate to the threat.

It follows that the potential negative consequences for individuals and the protection of fundamental rights due to the misapplication of counterterrorism powers (whether intentional or unintentional) do not justify the adoption of overly broad and vague definitions of terrorism. In the context of countering TVEC online, companies should "demonstrate the necessity and proportionality of any content actions (such as removals of account suspensions)" and in doing so consideration should be given to less intrusive restrictions (e.g., content warnings and de-amplification scaled to risk and degree of harm).[42]

Defining terms such as terrorism and violent extremism is not an easy task; however, tech companies have policies aimed at countering TVEC on their platforms and services. Consequently, they are required to explain to users what content constitutes TVEC and will therefore be removed.[43] Definitions will also help guide "the discretion of individual reviewers, reducing the risk of inconsistent – or even inappropriate – decision-making."[44] It is important that definitions are clear and unambiguous to avoid the possibility of "censorship creep,"[45] and to ensure compliance with human rights standards such as the principle of legality, necessity, and proportionality.

42 Report of the Special Rapporteur, A/HRC/38/35.

43 Stuart Macdonald, Sara Giro Correia, and Amy-Louise Watkin, "Regulating Terrorist Content on Social Media: automation and the rule of law," International Journal of Law in Context 15, no. 2 (2019): 183–197, https://doi:10.1017/S1744552319000119.

44 Macdonald, Correia, and Watkin, "Regulating Terrorist Content."

45 Danielle K. Citron, "Extremist Speech, Compelled Conformity, and Censorship Creep," Notre Dame Law Review 93, no. 3 (2018): 1035–1072.

Identifying Coherence and Divergence

An examination across the definitions revealed clear coherence and areas of divergence as to some core components/standard features of a definition of terrorism.

Definition of Terrorism Core Requirements: Coherence

Across the definitions, there was consensus that the definition of the term 'terrorism' includes the following core requirements:

- An act of violence;
- Indication of the level of harm resulting from the act;
- Proof of intention is necessary;
- The target is a wider audience beyond the immediate victims of the act; and
- Indication of the psychological impact on the target.

However, as is addressed below, within these core requirements there are layers of incoherence and inconsistency which is problematic when seeking to establish a common understanding of a definition of terrorism and the advantages of interoperability

Definition of Terrorism Core Requirements: Divergence

There exists divergence in relation to two core requirements of a definition of terrorism. These are in relation to the existence of a motive requirement and express exceptions.

Motive requirement

The motive requirement sets out reasons as to why an individual engages in the prohibited conduct, the underlying aim they sought to achieve, or the cause they sought to advance. For example, the requirement that the purpose of the conduct is to advance political, religious, racial, or ideological causes. The motive requirement is commonly understood to be the principal defining feature of terrorist attacks, distinguishing terrorism from ordinary crime.[46] Nevertheless, in the definitions of terrorism examined in this study, the existence of a motive requirement is identified as an area of divergence in the core requirements of a definition. Only five of the definitions explicitly include a motive requirement,[47] all of which include political causes.[48] The Australian and Meta definitions also include religious and ideological causes, while the Indonesian definition also includes ideological and security breach motivations. The UN Special Rapporteur's model definition of terrorism does not include a motive requirement, although it suggests that reference to motivations can "assist in further narrowing

46 Kent Roach, "The Case for Defining Terrorism with Restraint and without Reference to Political or Religious Motive," in Law and Liberty in the War on Terror, eds. Andrew Lynch, Edwina Macdonald, and George Williams (Sydney: Federation Press, 2007), 39; Bruce Hoffman, Inside Terrorism (New York: Columbia University Press, 2019), 38.

47 UN General Assembly Resolution 49/60 (December 9, 1994), US Title 22 USC § 2656f, and those for Australia, Indonesia, and Meta.

48 This is the sole motivation specified in two of those definitions: UN GA Res 49/60 and the U.S. (22 USC 2656f).

the scope of application of the definition of terrorism."[49]

Express exemptions

Definitions of terrorism often specify express exemptions. These include an express exemption for advocacy, protest, dissent, industrial action, and/or armed conflict. Across the definitions in this study, this is identified as an area of divergence, as only eight of the definitions include an express exemption. Of those, only Australia exempts advocacy, protest, dissent, and industrial action. The remaining include an armed conflict exception[50]– this is stated in all five of the regional instruments, the Terrorist Financing Convention, and Meta's definition.

If seeking a common understanding of the definition of terrorism, divergence on both the existence of a motive requirement and express exemptions raises important questions. On motive, there is the question of whether the motive requirement should be included in the definition, and if so, what motives should be included. Should the definition be restricted to political motives, or include a wider range such as religious and ideological motives? There is also the important question on the inclusion of express exceptions within a definition of terrorism for advocacy, protest, dissent, industrial action, and actions taken during armed conflict. The inclusion of such exemptions can assist in narrowing definitions of terrorism to ensure they comply with the principle of legality and therefore do not operate beyond their intended purpose, which in turn can ensure counterterrorism measures are necessary and proportionate to the threat. These issues are addressed in more detail in the "Implications of incoherence and divergence" section below.

49 Report of the Special Rapporteur, A/HRC/16/51, para 27.

50 Terrorist Financing Convention; EU Directive; OIC Convention; OAU Convention; Shanghai Convention; Arab Convention; Meta.

Divergence Within the Core Requirements of Definitions of Terrorism

While there is coherence as to the core requirements of a definition of terrorism identified in this study, there are layers of incoherence and inconsistency within each requirement, which is problematic when seeking to establish a common understanding and/or minimum standards. This is summarized in Table 1.

Table 1: Divergence Within the Core Requirements of Definitions of Terrorism

Core requirements: coherence	First layer of incoherence / inconsistency	Second layer of incoherence / inconsistency
Act of violence	Specific acts (listed)	
	General approach	Level of harm
		Existing criminal acts
	Combination of general and specific approach	
	Threats/attempts of action	
Indicating the level of harm resulting from the act	Range and level of harms	
Intention	To cause harm	Range and level of harms
	To impact the target	What the impact on the target is
	Both (cumulative)	As above
Target (wider audience)	Public/population	Level of impact
	Government	Level of impact
	International organization	Level of impact
	Public security/disorder	
Psychological impact on members of the public	Level of impact	

Act of Violence

Each of the definitions include an act of violence as a core requirement. However, there is divergence across the definitions as to how the act of violence is defined. Some of the definitions take a specific approach, setting out several specified acts that fall within the definition.[51] Alternatively, a general approach is taken to defining acts of violence either by reference to the level of harm caused (e.g., causing death or serious bodily injury),[52] by reference to existing criminal acts in national or international law,[53] or both.[54] Some definitions include a

51 Nuclear Terrorism Convention; EU Directive; U.S. (18 USC 2332b(g)(5)); France.

52 Shanghai Convention; Australia; U.S. (22 USC 2656f); U.S. (31 CFR 594); Indonesia; Meta.

53 UN GA Res 49/60; UN SC Res 1566; U.S. (18 USC 2331(1)); U.S. (18 USC 2332b(g)(5)).

54 Terrorist Financing Convention; OIC Convention; OAU Convention; Arab Convention; Special Rapporteur.

combination of the two approaches, for example Australia taking a general approach and including reference to some specific acts.

Divergence also exists as to whether the definition includes a threat of action[55] and/or attempted action. The UN Special Rapporteur has urged caution in this respect to ensure compliance with the requirements of the principle of legality.[56] Meta's definition goes further than threats and attempts and includes engaging, advocating, or lending substantial support to "purposive and planned acts of violence"; and the Shanghai Convention includes organizing, planning, aiding, and abetting. It is submitted that such facilitative actions should fall outside the definition of terrorism because by definition accomplices are not the perpetrators of the actual terrorist act. As with threats of action, these would be better dealt with in policies that address the circumstances in which posts and content about a terrorist act may be removed. It is suggested that definitions of terrorism ought to be confined to acts or attempted acts of violence.

Arguably, the specific approach of setting out a list of certain activities which constitute terrorist acts, without defining a general category of terrorism, better complies with the principle of legality.[57] Providing a list of specific acts such as hostage-taking and attacks upon a person's life can make it clearer and more certain what acts do and do not amount to terrorist acts.[58] This means that each act listed must be clear and precise, and within the definitions subject to this study the certainty of some terms can be called into question. For example, both the EU and French definitions include attacks upon the "physical integrity" of the person—further guidelines or examples of conduct would be helpful here.

An additional advantage of the specific approach is that it, "avoids political conflict over basic definitional principles," which in turn permits, "textual agreement to be reached."[59] Consequently, when considering the interoperability of definitions of terrorism, this approach could be seen as beneficial. Nevertheless, listing specific individual acts of terrorism, "might not be capturing what we mean by terrorism," because the additional requirements that distinguish a terrorist act from ordinary criminal acts may not be included. [60] Terrorism is not "inherent to any particular act or type of violence."[61] Another issue is that a specific approach may not be able to keep pace with new terrorist acts, particularly in relation to the use and development of new technology.[62] The combination of a general approach to defining additional core requirements of a definition of terrorism and listing specific acts, such as in the EU Directive and the French definition (which also include intent and purpose requirements), would address the first concern but not the latter.

55 Eight out of the 21 definitions included a threat of action across international, regional, national, and local levels: Nuclear Terrorism Convention; EU Directive; OIC Convention; OAU Convention; Arab Convention; Australia; U.S. (18 USC 2332b(g)(5)); Indonesia.

56 Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, "Australia: Study on Human Rights Compliance While Countering Terrorism," (A/HRC/4/26/Add.3, 14 December 2006).

57 Clive Walker, "The Legal Definition of 'Terrorism' in United Kingdom Law and Beyond," Public Law (2007): 331–352.

58 Martin Scheinin, "A Proposal for a Kantian Definition of Terrorism: Leading the World Requires Cosmopolitan Ethos," EUI Working Paper LAW, 2020/15. Scheinin identifies the EU Directive's approach of setting out an exhaustive list of enumerated pre-existing crimes as an effort to seek to preserve legality.

59 Geoffrey Levitt, "Is 'Terrorism' Worth Defining," Ohio Northern University Law Review 13, no. 1 (1986): 97.

60 Ben Golder and George Williams, "What is 'Terrorism'? Problems of Legal Definition," University of New South Wales Law Journal 27, no. 2 (2004): 270–295.

61 Anthony Richards, "Conceptualizing Terrorism," Studies in Conflict and Terrorism, 37:3 (2014): 213–236, https://doi.org/10.1080/1057610X.2014.872023.

62 Richards, "Conceptualizing Terrorism."

The general definitions of terrorism—defining terrorism with reference to certain criteria such as intention, and motivation—also include acts of violence. While there are benefits to the general approach to defining terrorism, such as flexibility and adaptability to the terrorist threat, this can pose difficulties for compliance with the principle of legality (both in ensuring acts of violence are defined clearly and precisely and that these are narrow enough to ensure the definitions are not used beyond their intended purposes). Arguably, the most coherent with the principle of legality are the definitions that require that **there can only be an act of terrorism if the act of violence corresponds to an existing criminal offence enacted for the purpose of compliance with an existing treaty against terrorism**.[63] All three of the general definitions at the international level and three out of the four general definitions at the regional level[64] seek to comply with the principle of legality by taking this approach. However, at the national level, only the U.S. definition of acts of terrorism[65] and the French definition do. The tech companies do not take this approach. The remaining general definitions of terrorism subject to this study define the act of violence with reference to the level of harm.

The level and range of harm resulting from the act of violence

There is consistency across the definitions that the terrorist act is violent and that this involves causing death or serious bodily injury. This does not pose particular difficulties from the perspective of the principle of legality. It follows that defining terrorism with reference to this higher level and narrow range of harms assists in ensuring definitions are sufficiently narrow to not result in a disproportionate response to the terrorist threat. Only the Terrorist Financing Convention restricts the range and level of harm in this way. Both the Special Rapporteur's model definition and UN SC Res 1566 go beyond this to include hostage-taking, which again is arguably clear and precise and does not result in an overly broad definition (when the other core requirements of a definition of terrorism are also met).

However, beyond this, there is a disparity in the range and level of harms that are required of a terrorist act. This adds a further layer of complexity in defining the core requirement of an act of violence within a definition of terrorism and identifying the level of intention where this is stated with reference to the harm caused by the primary act. The incoherence in the level and range of harms reveals a wide range of less serious and uncertain harms. For example, Meta's reference to "serious harm" may not be considered a clear and precise term, and circular definitions can cause difficulty in practice.
Several of the definitions include property damage,[66] but there are differences in the appropriate threshold and a lack of consistency as to whether property damage is linked to a risk of death or other injury or harm to individuals (e.g., endangering property; damage to property; exposing property to hazards; occupying or seizing property; substantial damage to property; serious damage to property; significant damage to property with a risk of death, serious injury or death; damaging or destroying property with a substantial risk of bodily injury; extensive destruction to property likely to endanger life or major economic loss). Similar divergences are evident in the definitions that include environmental damage.[67] There is a similar lack of coherence as to the appropriate threshold (e.g., exposing the environment to hazards; may cause damage to the environment; damage to the

63 Martin Scheinin, "A Proposal for a Kantian Definition."

64 The exception is the Shanghai Convention.

65 U.S. (18 USC 2332b(g)(5)).

66 Nuclear Terrorism Convention; EU Directive; Arab convention; OIC Convention; OAU Convention; Australia; U.S. (18 USC 2332b(g)(5)); U.S. (31 CFR 594); Meta.

67 Arab Convention; OIC Convention; OAU Convention; Indonesia.

environment; substantial damage to the environment).

The inclusion of property and/or environmental damage is potentially problematic from the perspective of legality, particularly when this is not linked to a risk of death or serious injury. This can result in overly broad definitions which can operate beyond their intended purposes, encompassing conduct that is not truly terrorist in nature. This could include political protest, for example, which results in or even deliberately causes serious property damage.[68] This problem is exacerbated in the absence of express exemptions and can lead to counterterrorism measures that are not necessary and proportionate to the terrorist threat.

Interestingly, only the Australian definition includes an action that creates a serious risk to public health or safety. Some of the definitions include impacts on natural resources,[69] again with varying thresholds of the level of harm (e.g., endangering; jeopardizing; damaging). The definitions that include acts of terrorism on infrastructure, systems, and facilities[70] also differ on the appropriate threshold of harm (e.g., seriously disrupts or destroys; damage or destruction; endangering; extensive destruction likely to endanger life or result in major economic loss). Issues of clarity and precision are evident in the inclusion of this range of harms, which in turn can lead to overly broad definitions of terrorism.

This raises several questions when seeking to establish minimum standards, including whether property and environmental damage should constitute terrorist acts—and if they do what is the appropriate threshold, and should this be linked to the risk of a loss of life or serious bodily injury? The same question arises as to whether attacks on infrastructure including facilities and systems (such as electronic systems) should be included—and if so what the appropriate level of harm should be. In addition, the EU Directive raises the question as to whether property damage and attacks on infrastructure should be linked to a risk of major economic loss. The alternative is to restrict an act of terrorism to one which causes death, endangers life, or causes serious bodily injury to members of the population.

To ensure coherence with human rights standards such as the principle of legality, necessity, and proportionality, it is proposed that the level and range of harms should be restricted to those that cause or endanger life, cause serious bodily injury, or involve hostage-taking or kidnapping. This is in conjunction with defining the act of violence as corresponding with a crime enacted for the purpose of compliance with an existing treaty against terrorism. This would help to preserve the principle of legality in definitions of terrorism, which in turn assists in ensuring counterterrorism measures are necessary and proportionate.

Target: Wider Audience

It has been argued that the "essence of terrorism lies in the intent or purpose behind the act of violence rather than in the act itself, namely, to generate a wider psychological impact beyond the immediate victims."[71] This

68 Kent Roach, "Sources and Trends in Post 9/11 Anti-Terrorism Laws," in Security and Human Rights, eds Benjamin J. Goold and Liora Lazarus (Hart Publishing, 2007); Jacqueline Hodgson and Victor Tadros, "The Impossibility of Defining Terrorism," New Criminal Law Review, 16, no. 3 (2013): 494–526.

69 Arab Convention; OIC Convention; OAU Convention.

70 EU Directive; Arab Convention; OIC Convention; Shanghai Convention; U.S. (31 CFR 594); Indonesia.

71 Anthony Richards, "Conceptualizing Terrorism." See also: A. Schmid and A. Longman, Political Terrorism, Third Edition (New Brunswick, NJ: Transaction Books, 2008); Hoffman, Inside Terrorism.

is captured across the definitions subject to this study, all of which refer to a target being a wider audience beyond the immediate victims of the act. There is coherence across the definitions that this includes the public, section of the public, or population.[72] The majority of the definitions also include the government,[73] and seven of the definitions also include international organizations.[74] These targets are clear and precise and correspond with the Special Rapporteur's model definition.

In addition, some of the definitions include public security (Shanghai Convention), public disorder (France), or public facilities (Indonesia) as targets at which terrorist acts can be directed. The EU Directive includes fundamental political, constitutional, economic, or social structures of a country or international organization. This presents issues with the principle of legality. These terms are not defined with precision, and even if the meaning is clear they present the danger of broadening a definition beyond its intended scope. This in turn could result in disproportionate counterterrorism measures.

The Psychological Impact on the Target

While specifying the target of a terrorist act as the population, public (or section of the public), the government, or international organization meets the requirements of legality, some of the definitions pose difficulties in how the psychological impact on the target is defined. The impact on the wider audience should be set at an appropriate threshold. This is particularly important as, for many of the definitions, this also sets the level of intention required for an act to be deemed an act of terrorism.

While coherence exists across the definitions of terrorism by including reference to the psychological impact on the target, there is divergence as to the appropriate threshold. Where the impact is the population or members of the public, the impact ranges (e.g., to coerce; to compel; to force; to instill fear; to intimidate; to seriously intimidate; to provoke a state of terror; to sow panic; to influence). Some definitions use a combination of these thresholds. In addition, where the target of the terrorist act includes the government, the threshold as to the level of impact differs (e.g., to coerce; to compel; to unduly compel; to force; to induce; to instill fear; to intimidate; to seriously intimidate; to seriously destabilize or destroy; to threaten stability; to influence; to affect; to retaliate against).

In this respect definitions that refer to influencing the government or the public arguably set the bar too low. This can lead to broad definitions, encompassing acts that do not constitute terrorist acts, such as protests and demonstrations designed to influence the government where some violence occurs. Similar difficulties arise with the use of terms such as 'affect,' which also could be deemed to lack clarity and precision. Some of the definitions also refer to instilling fear or sowing panic in the population. It has been previously noted in relation to the threshold of instilling fear that such circumstances could "result from non-political hooliganism or individual acts of aggression."[75] It has more recently been suggested in relation to proposed changes to the definition of terrorism in New Zealand that "a patched gang member … can induce fear at their local supermarket but it's

72 The exception here is France which refers to the impact on public disorder.

73 The exceptions here are the UN GA Res 49/60, UN SC Res 1566; Arab Convention; France, and Indonesia.

74 Terrorist Financing Convention; Nuclear Terrorism Convention; EU Directive; OAU Convention; Shanghai Convention; Meta; Special Rapporteur.

75 Clive Walker, "The Legal Definition of 'terrorism' in United Kingdom Law and Beyond."

arguable whether this qualifies as terrorism."[76] Therefore, using the lower threshold of "fear" could potentially encompass conduct that should be dealt with under the ordinary criminal law, which presents the danger of the definition operating beyond its intended purposes.

Setting the threshold of impact as intimidate, compel, or coerce would be more appropriate as the terms have a clearer meaning, providing better compliance with the principle of legality. Moreover, these terms are more purposive in nature rather than being an impact incidental to the act, and so also encapsulate the core concept of terrorism more accurately. The use of these terms would be appropriate in reference to the impact on both the population and governments and international organizations.

Intention

There is a consistency across all the definitions that an act of terrorism is planned, premeditated, or purposive. This is identified using wording such as "intention," "calculated," or "acts as part of an agenda." The requirement of intention is an area of coherence in the definitions of terrorism. Definitions can include the intentional primary act causing a certain level of harm (general intention), a specific or qualified intention to accomplish the purpose of impacting the target of the wider audience (whether this is the government or the population)[77] or require both a general and specific intention.[78] The Australian definition expressed the requirement as the intention to advance a political, religious, or ideological cause (motive), and the specific intent to accomplish the purpose of impacting the target. It should be noted that both the Arab Convention and the OIC Convention state that a terrorist act is carried out irrespective of the motives, intentions, and purposes behind the act; however, both go on to refer to seeking to sow panic or the aim of terrorizing people and causing fear. Therefore, an act of terrorism is purposive, and it appears that it requires both a general and/or specific intention, but this needs further clarity in the drafting.

Many of the definitions that refer to both general and specific intentions state these as in the alternative, with the use of the word 'or.' The Special Rapporteur supports a cumulative approach to defining terrorism, like SC Resolution 1566 (2004), including an intention to cause harm and an intention to impact the target.[79] This would better comply with the principle of legality, assisting in restricting definitions to actions that constitute terrorist acts.

Beyond whether the definitions of terrorism require a general or specific intention (or both), there exists an additional layer of incoherence. Where the definitions require a general intention to carry out the primary act of violence causing harm, as set out above the level and range of harms differ across the definitions. Similarly, in definitions that require a specific intention to accomplish the purpose of impacting the target, the threshold qualifying the level of intention differs (e.g., to coerce; to compel; to unduly compel; to force; to induce; to intimidate; to seriously intimidate; to influence; to affect; to instill fear; terror; seriously destabilizing or destroying).

76 Hayden Crosby, "Treating NZ's far right groups as terrorist organisations could make monitoring extremists even harder," The Conversation, 16 April 2021, https://theconversation.com/treating-nzs-far-right-groups-as-terrorist-organisations-could-make-monitoring-extremists-even-harder-158291.

77 UN GA Res 49/60 (9 December 1994); Australia; U.S. (18 USC 2331(1)); U.S. (18 USC 2332b(g)(5)); U.S. (31 CFR 594); France; Meta.

78 Terrorist Financing Convention; Nuclear Terrorism Convention; UN SC Res 1566; EU Directive; OAU Convention; Shanghai Convention; U.S. (22 USC 2656f); Indonesia; UN Special Rapporteur.

79 Report of the Special Rapporteur, A/HRC/16/51.

As discussed, above, many of the definitions use a different combination of these levels of intent. Therefore the issues identified with preserving legality with reference to the range and level of harm and in stating the psychological impact on the target of a terrorist act would need to be resolved to comply with human rights standards.

It should also be noted here that establishing intent can be difficult. The Terrorist Financing Convention attempts to provide some guidance, stating that intent to impact the wider audience is established "by its nature and context." Evidentiary issues are recognized specifically in establishing the specific or qualified intention, as this goes to the purpose of the perpetrator. This can be difficult to apply in practice and may be "inferred rather than proven."[80]

TVEC content moderation policies risk an unjustified interference with the freedom of expression. In seeking a common understanding and greater interoperability on definitions of terrorism that apply to prevent forms of expression online (as opposed to an act of violence), there is a risk of a chilling effect. Therefore a higher threshold of intention and a tighter definition of terrorism as it applies to identifying terrorist content is important.[81] It is submitted that **definitions should take a cumulative approach to intention, including an intentional primary act (of violence) and a specific intention to accomplish the purpose of impacting the target.** To maintain appropriately high thresholds, the general intention should be to cause death/endanger life and/or serious bodily injury and the specific intention should be qualified as to intimidate, coerce, or compel the population, government, or international organization.

Summary

The review of the definitions demonstrates that a minimum degree of coherence in the core requirements of terrorism exists. An act of terrorism involves an act of violence, carried out intentionally, with the purpose of impacting a specified target which includes members of the general population. However, layers of incoherence are evident in the level and range of harms resulting from the act of violence, the level of intent required, the range of targets, and the impact of the act of violence on those targets. There is also divergence evident across the definitions as to the existence of a motive requirement, what motivations that should include, the existence of express exemptions such as for protest, advocacy, industrial action, and dissent, and an armed conflict/IHL exemption.

Definitions of Violent Extremism

The Shanghai Convention is the only statutory instrument subject to this study that includes a definition of violent extremism. Australia and the U.S. provide no legal definition, although the term is defined in relevant policy documents which have been included in this report. Twitter provides a definition of violent extremism and YouTube's Community Guidelines state that content produced by violent extremist groups not included on terrorist designation lists is covered by its policies against posting hateful or violent content.

80 Martin Scheinin, "A Proposal for a Kantian Definition."

81 Alan Greene, "Defining Terrorism: One Size Fits All?," International and Comparative Law Quarterly, 66 (April 2017): 441–440. https://doi.org/10.1017/S0020589317000070.

Coherence is evident between these definitions that violent extremism involves acts of violence. However, beyond this, no further detail or clarification is provided as to the level or potential range of harms. This has the potential to be too broad, as it does not limit the threshold to serious violence. Twitter and the Shanghai Convention include as a requirement a target beyond immediate victims; however, there is divergence as the former refers to the population and the latter the government and public security. The U.S. and Australian policy documents include a motive requirement including political, religious, and ideological purposes.

There is considerable incoherence as all definitions are arguably vague and broad, presenting the danger that definitions will include individuals beyond the intended scope of countering violent extremism policies such as members of civil society.[82] The Special Rapporteur on counterterrorism and human rights has raised concerns about policies developed on the basis of the term, arguing that it is "conceptually weaker than the term terrorism, which has an identifiable core."[83] This raises the important question about companies developing policies aimed at countering violent extremist content in addition to terrorist content. Further research would be welcome as to whether content currently removed as violent extremist content would also be removed as terrorist content, or under existing policies for hateful or violent content.

Implications of Incoherence and Divergence

The Core Requirements of a Definition of Terrorism

While this paper has identified some standard features/core requirements of a definition of terrorism, **the inconsistency and incoherence within these requirements are problematic from the perspective of reaching a common understanding and increasing interoperability. This can lead to actions and content classified as terrorist in one jurisdiction or on one platform but not another.** Previous research has identified the example of where "a group carried out attacks against infrastructure without the intent to harm civilians (by releasing warnings)."[84] This would fall within some of the definitions subject to this study but not all. Some companies could judge this as not meeting its criteria for TVEC, but other companies may opt for removal. Similar outcomes may arise if a politically motivated group or individual carries out violence resulting in property damage but does not intend to cause death, endanger life, or serious injury to civilians, again falling within some definitions of terrorism but not all. Increased pressure to remove TVEC may result in some companies erring on the side of caution in these circumstances which can lead to over-censorship. Therefore, increasing coherence between definitions would be beneficial.

Existence of Motive Requirement

While the motive requirement is commonly understood to be the principal defining feature of terrorist attacks, distinguishing terrorism from ordinary crime,[85] it has been identified as an area of divergence in this study. The

82 Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, "Impact of measures to address terrorism and violent extremism on civic space and the rights of civil society actors and human rights defenders" (A/HRC/40/52, 1 March 2018).

83 A/HRC/40/52.

84 Isabelle van der Vegt, Paul Gill, Stuart Macdonald, and Bennett Kleinberg, "Shedding Light on Terrorist and Extremist Content Removal," Global Research Network on Terrorism and Technology: Paper No 3.

85 Roach, "The Case for Defining Terrorism," 39; Hoffman, Inside Terrorism, 38.

motive requirement differs from the intention requirement in that it is "directed towards the emotional reasons why the accused engaged in the prohibited conduct, as opposed to his or her desire to bring about a particular set of consequences."[86] The motive requirement goes beyond requiring proof that the perpetrator intended to commit the act itself and requires proof that the perpetrator engaged in the conduct for a particular reason (for example, in advance of a political cause). The Special Rapporteur does not include this in the model definition, although has stated that reference to motivations can "assist in further narrowing the scope of application of the definition of terrorism."[87]

Nevertheless, the inclusion of a motive requirement in a definition of terrorism has proved controversial. On the one hand, requiring proof of motive can narrow the definition of terrorism and reduce the risk of overly broad and disproportionate application of counterterrorism measures. On the other hand, the motive requirement is difficult to prove, and it is argued can lead to religious or racial profiling and can have a chilling effect on rights protection.[88]

Ben Saul has argued that the inclusion of motive distinguishes the "public-orientated reason" (such as religion) for carrying out violence constituting terrorism with "violence perpetrated for private ends."[89] He claims that definitions that only include an intention to target a wider audience, such as intimidating a population, do not accurately capture what is meant by terrorism. He points to extortion as an example of an action that intimidates a government or population which would be captured, albeit an act carried out for "private, non-political reasons"[90]—thereby leading to definitions of terrorism that extend beyond conduct that is truly terrorist in nature.

In addition, Saul argues that the inclusion of a motive requirement can enhance the protection of human rights by narrowing the definition. Referring to the U.K.'s definition of terrorism which includes racial motives, he proposes that the motive requirement can be "harnessed to confront and quell those who would terrorize others in the pursuit of racial supremacy or eugenic-fantasies."[91]

Specific Concerns with Religious Cause

Alternatively, critics of the inclusion of the motive requirement argue that it risks religious and racial profiling, prejudice, and a chilling impact on speech.[92] Concerns are raised specifically on the inclusion of religious

86 Kieran Hardy, "Hijacking Public Discourse: Religious Motive in the Australian Definition of a Terrorist Act," University of New South Wales Law 34, no. 1 (2011): 333–350.

87 Report of the Special Rapporteur, A/HRC/16/51. See also UN Office on Drugs and Crime, "Legislative Guide to the Universal Legal Regime Against Terrorism," 2008.

88 Roach, "The Case for Defining Terrorism."

89 Ben Saul, "The Curious Element of Motive in Definitions of Terrorism: Essential Ingredient or Criminalising Thought?" in Law and Liberty in the War on Terror, eds. Andrew Lynch, Edwina Macdonald, and George Williams, (Federation Press, 2007), 28.

90 Saul, "The Curious Element of Motive."

91 Saul, "The Curious Element of Motive."

92 per Rutherford J in the Canadian case R v Khawaja [2006] OJ 4245. Note higher courts did not agree: Court of Appeal for Ontario, R v. Khawaja, File Nos C50298–C50299, Neutral Citation No. 2010 ONCA 862, 17 December 2010; Supreme Court of Canada, R v. Khawaja, Case No. 34103, Neutral Citation No. 2012 SCC 69, 14 December 2012.

motive in a definition of terrorism.[93] The inclusion of a religious motive can directly fuel the misconception that Islamic religiosity causes terrorism or that the Qur'an generally radicalizes Muslims.[94] These are key narratives within extreme right movements:[95] in Australia, a prominent figure on the "radical right extremist scene" has been involved in "anti-Islam protest movements,"[96] and online links between these figures and the Christchurch attacker have been established.[97] Alternatively, terrorist groups such as al-Qaeda and Daesh have attempted to justify their actions based on a distorted rhetoric of Islam; by accepting their proclaimed identity as religious movements, it logically follows that public discourse will assume a link between Islamic jihad and terrorism.[98] Any suggestion of links between the actions of a terrorist organization and Islam is inappropriate and offensive to Muslim populations[99] and is considered an attack on the identity of the Muslim community, which adds to feelings of alienation.[100] Therefore, the inclusion of religious motives may be "misunderstood as targeting the entire group who wish to advance the religious cause of Islam,"[101] which is counterproductive.

These criticisms point to issues with the legality, proportionality, and necessity of this clause in achieving a legitimate aim. If a concept is so easily weaponized due to a lack of precision or clarity, it negates legality. If a clause undermines social cohesion and national security and contributes to more violent extremism, its proportionality and necessity also come into doubt.

Therefore, **it is recommended that religious causes should not be included in a motive requirement**.[102] "The requirement to prove religious motive in terrorism offenses comes too close to pursuing a case against a religion."[103] It is also arguably unnecessary, given that the objectives of groups such as al-Qaeda could also be described as political or ideological.[104]

Removing "religious cause" does not downplay the role of religious texts in violent extremist contexts, nor does it prevent people from being able to discuss religious texts and how they are used in violent extremism contexts. But it does reduce the likelihood of law or policy being used to further misconceptions and create

93 This is only included in the Australian definition in this study. This is modelled on the U.K.'s definition of terrorism (section 1 Terrorism Act 2000).

94 Anne Aly and Jason-Leigh Striegher, "Examining the Role of Religion in Radicalization to Violent Islamist Extremism," Studies in Conflict & Terrorism 35, no. 12 (2012): 849–862. https://doi.org/10.1080/1057610X.2012.720243; Faiza Patel, 'Rethinking Radicalization' (Research Report, Brennan Center for Justice at New York University School of Law, March 2011), 3.

95 Jade Hutchinson, "The New-Far-Right Movement in Australia," Terrorism and Political Violence, 33, no. 7 (2021): 1424–1446, https://doi.org/10.1080/09546553.2019.1629909.

96 William Allchorn, "Australian Radical Right Narratives and Counter Narratives in an Age of Terrorism," Hedayah and Centre for the Analysis of the Radical Right, 2021.

97 Allchorn, "Australian Radical Right Narratives."

98 Hardy, "Hijacking Public Discourse."

99 Hardy, "Hijacking Public Discourse." See also Victoria State Government, Expert Panel on Terrorism and Violent Extremism Prevention and Response Powers Report 2, 2017.

100 Victoria State Government, Expert Panel on Terrorism.

101 C. J. Gerard Brennan, "Liberty's threat from executive power," Sydney Morning Herald, July 6, 2007, https://www.smh.com.au/national/libertys-threat-from-executive-power-20070706-gdqjxj.html.

102 For further discussion, see Rita Jabri Markwell, "Religion as a Motive: Does Australian law serve justice?," Forum, forthcoming.

103 Australia Government Independent National Security Legislation Monitor, "Declassified Annual Report," December 20, 2021.

104 Hardy, "Hijacking Public Discourse." See also Mark Sedgwick, "Al-Qaeda and the Nature of Religious Terrorism," Terrorism and Political Violence 16, no. 4 (2004): 795–814, https://doi.org/10.1080/09546550590906098.

counterproductive outcomes. In the context of content moderation, the same test and considerations should apply. There are actors that spread demographic invasion theory about Muslims who thinly veil themselves as engaging in legitimate expression about "religiously motivated" terrorism and violent extremism. It is important that these veils are lifted and those actors properly investigated. Misconceptions and bias about Muslims and Islam among content reviewers and leadership must be actively resisted. Not highlighting religion within terrorism or violent extremism definitions is a clear way to do that. Platforms don't refer to white supremacy as "patriotically motivated violent extremism" because it would be giving undue credence to their self-proclaimed values. The same logic and approach should apply across the board.

Other Concerns with the Motive Requirement

Other segments of the community can be negatively affected by an overemphasis on motive rather than retaining a focus on the other elements of the crime.[105] Opponents to the inclusion of motivation in a definition of terrorism also raise pragmatic concerns such as evidentiary difficulties. Terrorism offenses do not sit easily within existing criminal law frameworks, the focus being on why the conduct was carried out (motive) and at who it was aimed (target).[106] The problem is that it can be difficult to determine what the particular purpose of a terrorist act was and why it was carried out. This can make prosecutions difficult. This will also be an issue for tech companies in developing content moderation practices in response to terrorist content online. In relation to the Australian definition, the difficulty identifying the perpetrator's motive was evident following the Lindt Café Siege in Sydney, which was generally accepted as a terrorist attack. The State Coroner of New South Wales concluded that,

> **Even with the benefit of expert evidence, it remains unclear whether Monis was motivated by IS to prosecute its bloodthirsty agenda or whether he used that organization's fearsome reputation to bolster his impact. Either way, he adopted extreme violence with a view to influencing government action and/or public opinion concerning Australia's involvement in armed conflict in the Middle east. That clearly brings his crimes within the accepted definition of terrorism.**[107]

However, without proof of the perpetrator's motivations, it is difficult to conclude that this attack meets the requirements of Australia's statutory definition of terrorism.[108] For this reason, the expert panel recommended removing the motive element and strengthening the intention requirement to include an intent to provoke or create a state of terror. Even prior to this, based on both principled and pragmatic concerns, Australia's Independent National Security Legislation Monitor had recommended removing the motive requirement from the definition in his 2012 Report.[109]

The inclusion of the requirement of motive was an anomaly in this study and only included in five of the

105 For example, autistic people have been historically overrepresented in Prevent program referrals in the UK; see Jamie Grierson, "'Staggeringly high' number of autistic people on UK Prevent scheme," The Guardian, July 7, 2021, https://www.theguardian.com/uk-news/2021/jul/07/staggeringly-high-number-of-people-with-autism-on-uk-prevent-scheme.

106 Bernadette McSherry, "Terrorism Offenses in the Criminal code: Broadening the Boundaries of Australian Criminal Laws," UNSW Law Journal 27, no. 2 (2004): 354–372.

107 State Coroner of New South Wales, "Inquest into the Deaths Arising from the Lindt Café Siege: Findings and Recommendations," May 2017.

108 Victoria State Government, "Expert Panel on Terrorism."

109 Australia Government, "Declassified Annual Report," Recommendation VI/1, December 20, 2012.

definitions of terrorism. Therefore, it is not recommended for inclusion in the interests of moving towards greater interoperability and identifying a common understanding. This is not to say that there is not a role for assessing the motivations in the broader analysis of the context of certain actions or material. This can be useful in "illuminating particular aspects of the phenomena of terrorism and of terrorists."[110] However, based on this study it would not be deemed to be a necessary component of a legal definition of terrorism. If the intention requirement is drafted with clarity and included in a definition of terrorism, these purposes are "inherently political or broadly social phenomena."[111] In other words, if the definition of terrorism conveys the intention to intimidate the target of a wider audience, it is doubtful whether it needs to go on to expressly include reference to motivations.

Existence of Exceptions

A general behavioral approach to defining terrorism also presents the danger of encompassing conduct which is not truly terrorist in nature. This shows the importance of ensuring that such conduct is excluded from a definition of terrorism.[112] The existence of express exclusions on the face of the definitions is arguably anomalous.[113] However, the existence of exceptions is not in itself anomalous in the context of applicable human rights standards and sources of international law such as IHL.

Only the Australian definition excludes protest, advocacy, dissent, and industrial action. Without such an exemption, if a political protest became violent, this would arguably fall within the definition of terrorism. At national level, such as in the U.S., strong constitutional protection is given to the right to freedom of speech under the First Amendment of the Bill of Rights. Arguably, this would ensure that definitions of terrorism would not be applied to protest. However, such overriding constitutional protection of the freedom of speech does not exist globally and would not apply to all tech companies countering terrorist content on their platforms. It is difficult for companies to differentiate between TVEC and protest and dissent. This is evident in the case of Kurdish activists in Turkey advocating for an independent Kurdistan. One group—the Kurdish Workers' Party (PKK)—is designated as a terrorist group by governments. However, Kurdish activists have previously alleged that Meta removed posts in breach of its community guidelines, where individuals were engaging in mere legitimate dissent.[114] Therefore, **it is recommended that definitions of terrorism should expressly exclude protest, dissent, and industrial action**, and that the existence of an exemption is a harmful area of divergence identified in this study.

The second (perhaps more controversial) exception to consider is that of terrorism in a just cause. If a definition of terrorism applies to actions committed anywhere in the world against any government, then liberation and resistance movements could fall within its scope. This is irrespective of how oppressive or undemocratic the government is. The difficulties in such an exception are encapsulated by the phrase, "one person's terrorist is

110 Additional Progress Report Prepared by Ms. Kalliopi K. Koufa, Special Rapporteur on Terrorism and Human Rights, Sub-Commission on the Promotion and the Protection of Human Rights, E/CN.4/Sub.2/2003/WP.1 13, August 8, 2003.

111 Australia Government, "Declassified Annual Report."

112 Golder and Williams, "What is 'Terrorism'?"

113 Eight of the definitions included express exemptions; of these only Australia includes an exception for protest, etc.

114 Abdul Rahman Al Jaloud et al., "Caught in the Net: The Impact of 'Extremist' Speech Regulations on Human Rights Content," Electronic Frontier Foundation, May 2019, https://www.eff.org/wp/caught-net-impact-extremist-speech-regulations-human-rights-content.

another's freedom fighter," and the view that "those who opt for terror always believe their cause is just."[115] One way to alleviate this issue is to limit the definition to violence against non-state actors, which is Meta's approach.[116] In addition, some of the definitions expressly exclude activities carried out during armed conflict as determined under international humanitarian law.[117]

While the inclusion of such an exemption would not eliminate all situations where the actions of liberation or resistance movements operating in a non-international armed conflict could be classified as terrorist, an exclusion clause can assist in regulating the relationship between counterterrorism law and international humanitarian law.[118] Otherwise, in conjunction with broad general definitions, the consequence could be that a definition of terrorism could "criminalize certain activities carried out overseas that constitute lawful hostilities under international law."[119] U.K. case law offers a useful illustration of the difficulty.[120] This resulted in a conviction for a terrorist-related offense for an individual posting videos on YouTube showing attacks on coalition forces in Iraq and Afghanistan.[121]

The lack of an exception, and not limiting definitions of terrorism to violence against non-state actors, could have "significant repercussions for activists living in oppressive regimes."[122] This is an important consideration in the context of moderating terrorist content online, particularly given the increasing role of social media in contemporary activism.[123] In Egypt, research has shown that social media helped activists coordinate their "collective, offline actions" during the Egyptian revolution that led to overthrowing the Mubarak regime.[124] With companies facing heavy penalties, increasing government pressure to remove TVEC[125] can lead to an adverse impact on "collective action efforts"[126] and risks over censorship, as tech companies may err on the side of caution (which increases the potential for mistaken classification). This can lead to the deletion of important information such as evidence of human rights violations or war crimes.[127] This was illustrated by

115 George Fletcher, "The indefinable concept of terrorism," Journal of International Criminal Justice 4 (2006): 894–911.

116 Note that Twitter also takes this approach in its definition of violent extremism. See also Chris Meserole and Daniel Byman, "Terrorism Definitions and Designation Lists: What Technology Companies Need to Know," Global Research Network on Terrorism and Technology: Paper No. 7, 2019.

117 See further Thomas Van Poecke, Frank Verbruggen, and Ward Yperman, "Terrorist offenses and international humanitarian law: The armed conflict exclusion clause," International Review of the Red Cross 103, (2021): 295–324, https://doi.org/10.1017/S1816383121000321.

118 Van Poecke, Verbruggen, and Yperman, "Terrorist offenses."

119 Independent Reviewer of Terrorism Legislation, "The Terrorism Acts in 2011," June, 2012.

120 While the UK's statutory definition of terrorism is outside the scope of this study, the definition lacks an express exemption.

121 R v Gul [2013] UKSC 64. For further discussion see Alan Green, "The Quest for a Satisfactory Definition of Terrorism: R v Gul," Modern Law Review 77, no. 5 (2014): 780–807, https://doi.org/10.1111/1468-2230.12090.

122 Macdonald, Correia, and Watkin, "Regulating Terrorist Content."

123 Macdonald, Correia, and Watkin, "Regulating Terrorist Content."

124 Kara Alaimo, "How the Facebook Arabic Page 'We Are All Khaled Said' Helped Promote the Egyptian Revolution," Social Media + Society, (July 2015): 1–10, https://doi.org/10.1177%2F2056305115604854. See also Paulo Gerbaudo, Tweets and the Streets: Social Media and Contemporary Activism (London: Pluto Press, 2012).

125 Tech Against Terrorism, "The Online Regulation Series: The Handbook," July 2021, https://www.techagainstterrorism.org/wp-content/uploads/2021/07/Tech-Against-Terrorism-%E2%80%93-The-Online-Regulation-Series-%E2%80%93-The-Handbook-2021.pdf.

126 William Lafi Youmans and Jillian C. York, "Social media and the activist toolkit: User agreements, corporate interests, and the information infrastructure of modern social movements," Journal of Communication 62, no. 2 (April 2021): 315–329, https://doi.org/10.1111/j.1460-2466.2012.01636.x.

127 Svea Windwehr and Jillian C. York, "The Invisible Content Cartel that Undermines the Freedom of Expression Online," VOX-Pol Blog, November 4, 2020, https://www.voxpol.eu/one-database-to-rule-them-all/.

YouTube's removal of thousands of videos documenting the civil war in Syria after it introduced new technology designed to "identify violent content that may be extremist propaganda or disturbing to viewers" in breach of the company's community guidelines.[128] In addition, Meta came under criticism for the removal of images documenting ethnic cleansing in Myanmar.[129] These difficulties persist, and in response to the Russian invasion of Ukraine, a joint civil society letter was sent to Google, Meta, Telegram, Tik Tok, and Twitter calling on companies to improve content moderation practices in crisis situations, which included providing clarity about how TVEC is defined.[130]

There are circumstances when designations of terrorism might not be the most appropriate or effective way of approaching the problem—where groups are involved in armed conflict and their activities are confined to that armed conflict. Under these circumstances, the targeting of civilians through terror should be condemned as violations of the Law of Armed Conflict and the Geneva Conventions. If there is a lawful basis for violence, describing those individuals or groups as terrorists compromises democratic resolution by distorting public discourse. For tech companies, having an exemption for armed conflict means that material breaching the law would be properly classified as war crime material but not terrorist in nature.

Consequently, it is recommended that definitions of terrorism expressly exclude activities carried out during armed conflict as determined under international humanitarian law. It would also be beneficial to limit definitions to violence carried out by non-state actors.

Final Thoughts and Next Steps

This paper argues that tech companies should define terrorism and not rely on list-based approaches. There is an advantage for GIFCT member companies to move towards greater interoperability. However, in seeking a common understanding it is important to avoid arriving at the lowest common denominator in a definition that is not compliant with human rights standards.

Definitions should comply with the principle of legality in that they are clear and unambiguous and do not allow for counterterrorism measures to extend beyond their intended scope. This in turn will assist in ensuring that measures taken to combat terrorist content online are necessary and proportionate.

This paper has identified a minimum degree of coherence in the core requirements/standard features of a definition of terrorism. An act of terrorism involves an act of violence, carried out intentionally, with the purpose of impacting a specified target which includes members of the general population. However, it has identified many inconsistencies within these core requirements which are problematic, as actions and content can be classified as terrorist in one jurisdiction or on one platform but not another.

With this in mind, this paper recommends minimum standards in the core requirements of a definition of

128 Avi Asher-Schapiro, "YouTube and Facebook are Removing Evidence of Atrocities, Jeopardizing Cases Against War Criminals," The Intercept, November 2, 2017, https://theintercept.com/2017/11/02/war-crimes-youtube-Meta-syria-rohingya/.

129 "Facebook bans Rohingya group's posts as minority faces 'ethnic cleansing'," The Guardian, September 20, 2017, https://www.theguardian.com/technology/2017/sep/20/Meta-rohingya-muslims-myanmar.

130 Dia Kayyali, "Mnemonic Joins Open Letter Calling on Social Media Platforms to Improve Practices Globally," Mnemonic (Blog), April 22, 2022, https://mnemonic.org/en/content-moderation/Mnemonic-open-letter-social-media-platforms.

terrorism. Acts of violence should constitute pre-existing criminal ssssssss, either enacted for the purpose of compliance with an existing treaty against terrorism or identified as a serious crime in national law. The range and level of harm caused by the act should be restricted to those that cause death/endanger life, cause serious bodily injury, or involve hostage-taking or kidnapping. The purpose of the act is to impact a target being a wider audience beyond the immediate victims, including the population, the government, or an international organization. Proof of intention is necessary, and definitions should take a cumulative approach to intention, including a general intentional primary act (of violence) and a specific intention to accomplish the purpose of impacting the target. The specific intention should be qualified as to intimidate, coerce, or compel.

It is recognized that general definitions of terrorism present the danger of applying to conduct that does not constitute a terrorist act. While applicable human rights standards can act as a safeguard to a certain degree, it is not the case that the same standard of rights protection applies across jurisdictions and to tech companies. Therefore, it is recommended that definitions include expressexemptions such as protest, industrial action, advocacy, and dissent, and exclude activities carried out during armed conflict as determined under international humanitarian law. Moreover, to avoid the application of TVEC moderation policies to individuals in oppressive regimes, it would be beneficial to limit definitions to violence against non-state actors.

Table 2: Minimum Standards in the Core Requirements of a Definition of Terrorism

Minimum Standards in the Core Requirements of a Definition of Terrorism	
Act of violence	Constituting pre-existing criminal offenses, enacted: 1. for the purpose of compliance with existing treaty against terrorism; or 2. a serious crime in national law
Indicating the level of harm resulting from the act	Cause death, endanger life, cause serious bodily injury, or involve hostage-taking or kidnapping
Target (wider audience)	Population, government, international organization
Psychological impact on target	Intimidate, coerce, or compel
Proof of intention is necessary	Cumulative approach: general intention to carry out the act of violence and a specific intention to accomplish the purpose of the psychological impact on the target
Expressly exclude	Protest, industrial action, advocacy, and dissent
Expressly exclude	Activities carried out during armed conflict as determined under international humanitarian law

The inclusion of a motive requirement generally was identified as an anomaly or outlier in the definitions included in this study, and it is recommended in the interests of moving toward greater interoperability and identifying a common understanding that definitions of terrorism should not include a religious motive.

A query was raised during deliberations on this paper about the purpose of violent extremism as a label and

category for platforms—specifically (1) whether the label encouraged the explanation of rights-impacting measures beyond what is envisaged by lawmakers and (2) whether more previse labels that have a clear basis in law were preferable (for example, incitement of violence or publication of violent material). Further research into the purpose and value of tech companies defining violent extremism is recommended. It would be useful if transparency reports included data illustrating how much content is removed as being violent extremism but not terrorist in nature and the procedures that lead to such a decision. It follows that it would be useful if tech companies more clearly and precisely define some of the related terminology used in their terms of service such as harmful, offensive, and violent. This would help to ensure that content related to violent extremism is suitably captured by content moderation policies.

The scope of this study has been to examine the interoperability of definitions of terrorism/terrorist acts. This is only a starting point and as next steps a further area of timely research would be to examine definitions of terrorist content with the view to identifying clear and workable definitions for tech companies. Particular attention should be directed to the difficult question of how to determine the intention of the user posting particular types of TVEC,[131] as the scope has been limited to the intention of an individual committing a terrorist act. The circumstances in which online posts about a terrorist act are a separate area of inquiry.

Recommendations

- Tech companies should clearly define terrorism and not rely solely on list-based approaches. It follows that there is an advantage for tech companies in moving towards greater interoperability with respect to definitions of terrorism.
- There are minimum standards identified in this paper that should be included within the core requirements/standard features of a definition of terrorism.
- Definitions should expressly exclude protest, industrial action, advocacy, dissent, and activities carried out during armed conflict as determined under international humanitarian law.
- Religious motives should not be included in definitions of terrorism. A general motive requirement is not recommended in the interest of moving towards greater interoperability.
- Further research is needed into the value of companies defining violent extremism. This would include transparency in outcomes and procedures of TVEC removed as being violent extremist but not terrorist.
- Further research is needed to identify clear and workable definitions of terrorist content. Particular attention needs to be given to the issue of establishing the intent of a user.

Acknowledgments

This paper was authored by Katy Vaughan (Swansea University) on behalf of the definitions subgroup led by Rita Jabri Markwell (AMAN).

The study was designed and informed by members of the definition subgroup of GIFCT's LFWG. We would also like to thank Micalie Hunt (GIFCT) for her assistance at the initial stages of data collection.

131 On this issue in the context of the U.K.'s Draft Online Safety Bill, see Independent Reviewer of Terrorism Legislation, "Missing Pieces: A Note on Terrorism Legislation in the Online Safety Bill," April 20, 2022. https://terrorismlegislationreviewer.independent.gov.uk/missing-pieces-terrorism-legislation-and-the-online-safety-bill/.

Full List of Participating Individuals and Organizations

Rita Jabri Markwell (AMAN)	Jamie Brown (Council of Europe) (Counter-Terrorism Division)
Valère Ndior (University de Bretagne occidentale)	Anna Sherburn (Commonwealth Secretariat)
Department of Home Affairs (Australia)	Kabir Darshan Singh Choudhary (Wikimedia Foundation)
Dia Kayyali (Mnemonic)	Lauren Krapf and Jenna Hopkins (Anti-Defamation League)
Clare Allely (University of Salford)	Dima Samaro (Researcher)

Appendix 1

Definitions of Terrorism

International Instruments

International Convention for the Suppression of the Financing of Terrorism, opened for signature on 9 December 1999, No. 38349 (entered into force 10 April 2002)

International Convention for the Suppression of Acts of Nuclear Terrorism, opened for signature on 13 April 2005, No. 44004 (entered into force 7 July 2007)

Nuclear Terrorism Convention

UN GA Res 49/60 (9 December 1994)

UN GA Res 49/60

UN SC Res 1566 (2004).

UN SC Res 1566

Regional Instruments

Directive (EU) 2017/541 of the European Parliament and of the Council of 15 March 2017 on combating terrorism

EU Directive

Organization of Islamic Cooperation, Convention on Combating International Terrorism, Annex to Resolution No: 59/26-P (1 July 1999)

OIC Convention

Organization of African Unity, Convention on the Prevention and Combating of Terrorism, open for signature on 1 July 1999 (entered into force 6 December 2002)

OAU Convention

Shanghai Cooperation Organization, Shanghai Convention on Combating Terrorism, Separatism and Extremism, opened for signature 15 June 2001 (entered into force 29 March 2003).

Shanghai Convention

League of Arab States, The Arab Convention for the Suppression of Terrorism, opened for signature 22 April 1998 (entered into force 7 May 1999)

Arab Convention

Domestic Legislation

Australia, Section 100.1 of the Criminal Code.

U.S., Title 18 United States Code § 2331(1)

U.S. (18 USC 2331(1))

U.S., Title 18 United States Code § 2332b(g)(5)

U.S. (18 USC 2332b(g)(5))

U.S., Title 22 United States Code § 2656f

U.S. (22 USC 2656f)

U.S., Title 31 Code of Federal Regulations § 594

U.S. (31 CFR 594)

France, Article 421-1 (Code Penal) (Amended by LAW n° 2016-819 of June 21, 2016 – art. 1)

Indonesia, Act 5, 2018 Ch1 P2.

Tech Companies

- Meta
- Twitter
- YouTube
- Microsoft

Special Rapporteur

Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, "Ten areas of best practice in countering terrorism" (A/HRC/16/51, 22 December 2010).

Definitions of Violent Extremism

Regional

Shanghai Cooperation Organization, Shanghai Convention on Combating Terrorism, Separatism and Extremism, opened for signature 15 June 2001 (entered into force 29 March 2003).

Shanghai Convention

National

Australia

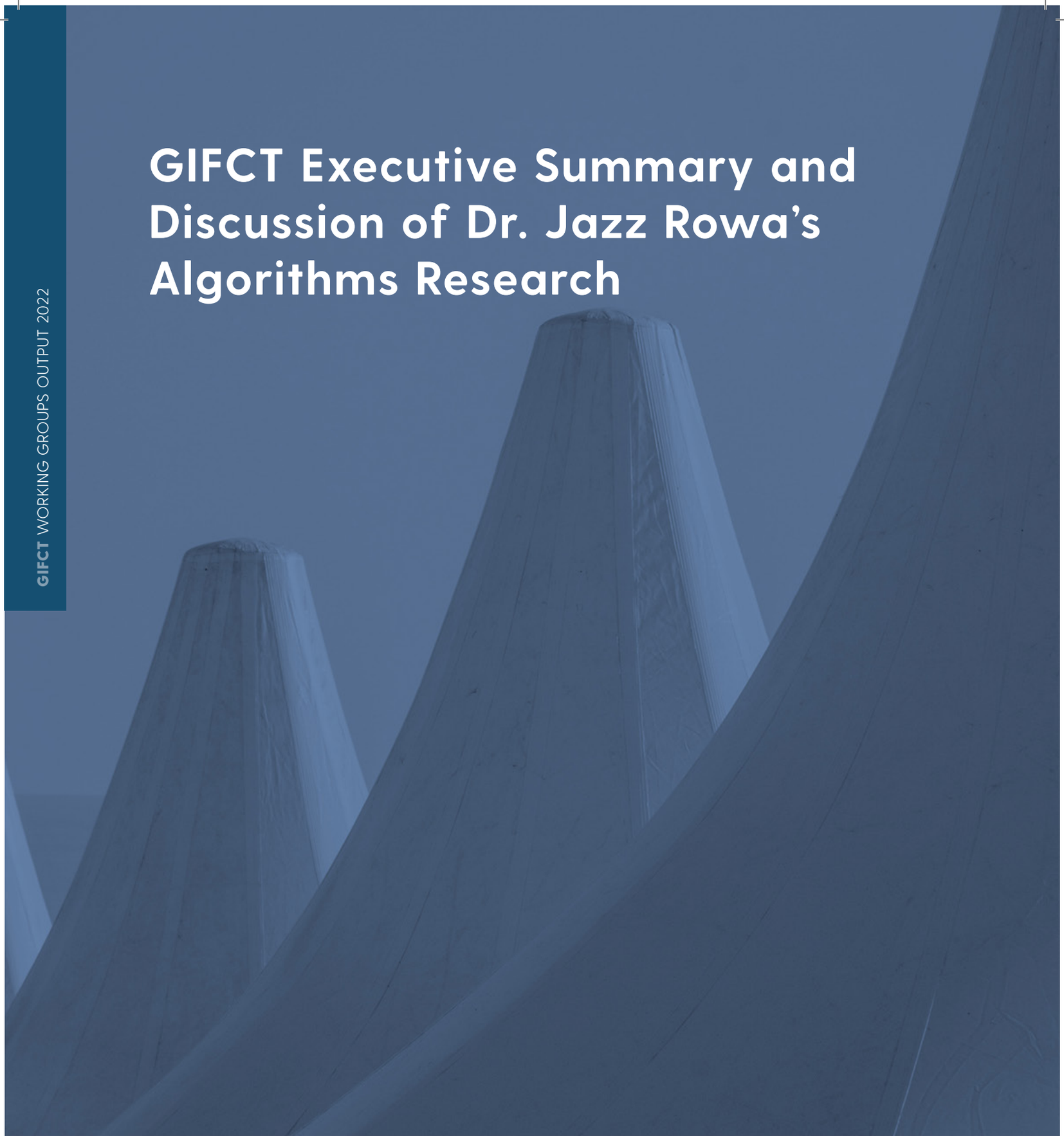
- Australian Government Department of Foreign Affairs and Trade, "Developing Approaches to Countering Violent Extremism."
- Australian Government Department of Home Affairs, "Countering Violent Extremism."

U.S.

- Strategic Implementation Plan for Empowering Local Partners to Prevent Violent Extremism in the United States.
- Government Accountability Office report, "Countering Violent Extremism: Actions Needed to Define Strategy and Assess Progress of Federal Efforts"

GIFCT WORKING GROUPS OUTPUT 2022

GIFCT Executive Summary and Discussion of Dr. Jazz Rowa's Algorithms Research



GIFCT
Global Internet Forum to Counter Terrorism

Dr. Erin Saltman
Director of Programming,
GIFCT

GIFCT recognizes the increasing concern from governments, researchers, technologists, and human rights advocates about the potential link between algorithmic amplification and processes of radicalization towards violence. Increased legislative language around the world has turned to 'algorithmic transparency' and one of the primary themes of the Christchurch Call to Action's Second Anniversary Summit in 2021 was to support methods to better understand user journeys online and the role algorithms may play in processes of radicalization. There is a fear that the nature of online environments may amplify hatred and glorify terrorism and violent extremism in a way that drives others towards violence. To effectively counter terrorism and violent extremism online, GIFCT aims to support research, analysis, and tools to better understand the true nature of the problem so that action can be taken. On the topic of understanding algorithmic processes there remain large knowledge gaps. GIFCT commissioned an extensive research effort by Dr. Jazz Rowa to assist in framing and better understanding the role of algorithms as part of GIFCT's 2022 Working Group outputs. This executive summary of her longer research paper, The Contextuality of Algorithms: An Examination of (Non)Violent Extremism in the Cyber-Physical Space, serves as a briefing document and reflection from GIFCT about some of Dr. Rowa's key findings. As of September 2022, her longer report can also be found on the GIFCT website under Working Group output and under our highlighted resources.

Background

In the first year of GIFCT Working Groups, held September 2020 through July 2021, GIFCT convened a group of global experts focussed on Content-Sharing Algorithms, Processes, and Positive Interventions, with participants from across tech companies, government, and civil society. Since an algorithm can be almost any input online with an output, the group adopted the shared goal of mapping which content-sharing algorithms and processes used by industry had the potential of facilitating consumption of content that may amplify terrorist and violent extremist content, or user interest in such content. The group also mapped and considered positive interventions and risk mitigation points for safety-by-design. The results of this paper honed in on the algorithmically optimized surfaces and tools that could potentially be exploited by bad actors, such as terrorists or violent extremists. This allowed the conversation on algorithms to focus more specifically on three online surfaces: search functions, recommendation features, and ad targeting algorithms.

In Year 2 of Working Groups, held September 2021 through July 2022, GIFCT commissioned Dr. Jazz Rowa to take this conversation and analysis further. GIFCT Working Groups had sub questions related to algorithms and the nexus with extremism in 3 of our 5 groups and asked Dr. Rowa to sit across these groups to develop this extensive paper. She has provided 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 participation was supplemented with empirical research and a range of first-person interviews. 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.

Findings

While this paper presents a myriad of findings and poses further questions, identifying gaps for further research,

there are some key takeaways that stuck out to our teams at GIFCT, which we will be processing and looking to build further work around in the future. The first takes us back to the age-old questions of definitions. In group discussions and interviews it remains clear that **there is no overarching agreement between different sectors or geographies on what online terrorist content is, what violent extremism is, what algorithms are, and what "extremist" or "borderline" content is.** If it can't be well defined, or if legislative language is vague on these points, we are still left with too much ambiguity to apply technical solutions or to ensure rigorous oversight or accountability mechanisms. Specifically for online spaces, the better you can define harm parameters the more you can measure, evaluate and risk mitigate. Vague or ambiguous terminology can lead to over censorship, under censorship, or the inability to measure and understand the nature of the problem in the first place.

While pressure escalates for tech companies to "do more", the analysis notes that **the current guidance on human rights in national, regional, and international legal frameworks is technologically suboptimal.** The pressure to expand technical solution-building is not equally matched with practical guidance of what human rights applications for technological ecosystems should look like. The paper also found that even some government representatives were wary that the term "algorithm" had become the latest buzzword and hot topic in the international debates on preventing and countering terrorism and violent extremism online, without enough clarity on the concept or the scope.

Dr. Rowa addresses the multiple reasons **why understanding algorithms, and attempts to provide meaningful algorithmic transparency, remains difficult. There is a notable difference between algorithmic explicability, interpretability, and auditability.** However, approaching algorithmic systems and its "black box" effect for analyzing input and output variables is compounded for a number of reasons; very few people understand the technical side of digital technologies, there remains a system of self-regulation for the technical evolution and review of technologies, there are methodological limitations for external researchers reviewing algorithmic systems, all combined with a trend of reactionary government regulation. The disclosure of an algorithmic formula or source code is viewed by some as useful and many as irrelevant in understanding a program's predictive behavior. Meanwhile there is a multi-dimensional and ever-changing landscape for both terrorist and violent extremist actors online and technical dynamism of platforms themselves. This conceptualisation of audits and the design of mechanisms for algorithmic oversight must therefore acknowledge the complexity of such an undertaking. To work towards greater algorithmic transparency, more work will need to be done to fully understand what "meaningful" data and algorithmic transparency means to policy makers and relevant stakeholders. Data and information sharing from tech companies can take many forms and alignment on understanding what data is useful and meaningful is crucial.

The current discourse on the role of algorithms in (non)violent extremism has for the most part created a false dichotomy between the online and offline spaces. The discussion around user, platform, and government furthers the complexity in trying to interpret causality in processes of radicalization and agency. User agency and lived experiences particularize contextual phenomena and inform the integration of the online and offline dynamics of extremism. Dr. Rowa points out that the interplay between the user and how an algorithm operates is intrinsically tied. Algorithmic systems are representations of human decisions and worldviews. What happens in the online realm cannot be detached from real life actions. This interplay needs to also inform legislative thinking.

Related to the discourse around user and platform accountability and responsibility, the interviews highlighted

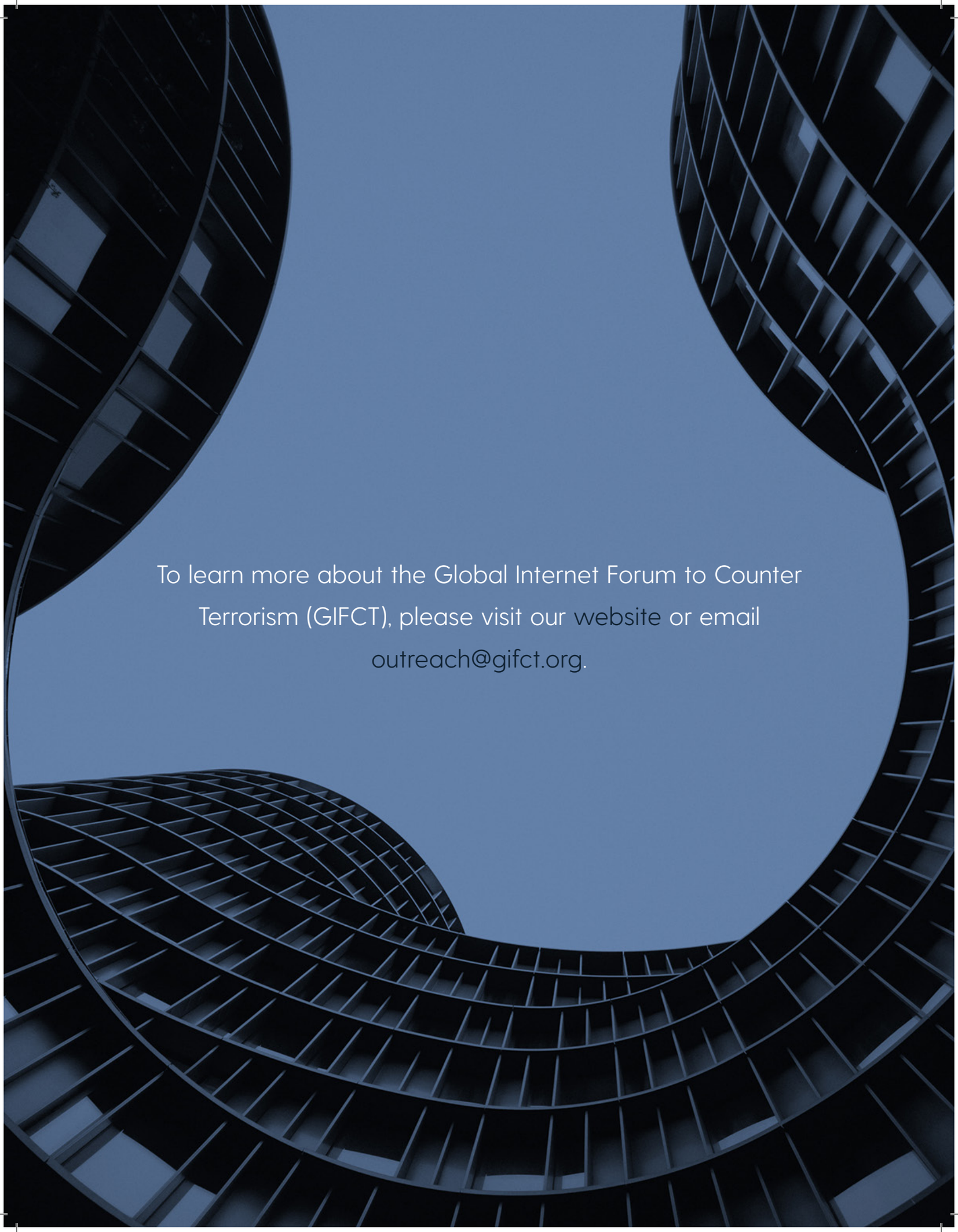
the **continued discomfort with non-violent and non-violating extremist content in what might be determined "gray area" content, and what, if anything, tech companies should do about it.** If users create legal, non-violating content and other users actively search and engage with the content, should private technology companies exert absolute control over the curation and restriction of legal but 'extreme' content? The concerns over borderline content are tied to the overarching debate on the definition of extremist content, liability for content creation, and the dispersal of content across digital publics (within hybridized or algorithmically amplified systems).

While some algorithm/user interplay could potentially amplify extremist content, there remain many spaces online that are beacons for violent extremist and terrorist sympathizers, yet have no algorithmic optimization associated with content surfacing or group recommendation features. These platforms remain a beacon to hate-based groups simply because they lack proactive moderation of content. The analysis notes that the recent lone actor attack in Buffalo, New York is seen as a case of "radicalization on 4chan" by other users giving social constructive information, documents, and social feedback. The attacker was also previously known to police, meaning there were offline signals that could have been used to provide support or have led to PVE/CVE interventions.

The overall research creates many avenues for further dialogues and multistakeholder work. However, it is important to recognize where positive opportunities for future work lie. **The research concludes that algorithmic processes, while being the core scrutiny of this paper, are equally where solutions can be found.** Despite the initial research question for the paper, Dr. Rowa points out that, paradoxically, algorithmic systems are conceived as automated problem solvers. In concert with other agencies, algorithms can act as conduits for the reconciliation, remediation, and reconstitution of an increasingly dysfunctional cyber-physical order. Whereas algorithms pose (un)known challenges for extremism, the opportunities they present in the mitigation and resolution of this and other societal challenges is equally consequential.

We at GIFCT hope that this research is of utility to the broadest range of stakeholders working to counter terrorism and violent extremism online and are grateful to Dr. Jazz Rowa for the time and energy she put into this extensive research over the last year.

Dr. Erin Saltman
Director of Programming
GIFCT



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