GIFCT
Global Internet Forum
to Counter Terrorism

Broadening the GIFCT Hash-Sharing Database Taxonomy: An Assessment and Recommended Next Steps

July 2021

GIFCT

Global Internet Forum to Counter Terrorism

Acknowledgements

The Global Internet Forum to Counter Terrorism would like to thank all of our stakeholders and partners for their interest in and commitment to this body of work. We are especially grateful to our GIFCT member companies, members of our Independent Advisory Committee (IAC), and civil society representatives for their feedback throughout this process, as well as to the GIFCT interns—Maggie Frankel, Aaron Tielemans, and Ye Bin Won—who provided significant support in the production of this series. GIFCT is also enormously appreciative of the time, expertise, and innovative thinking that all of the authors of the subsequent papers contributed to this project.

Table of Contents

GIFCT
Global Internet Forum
to Counter Terrorism

Introduction

By Nicholas J. Rasmussen and Johannah Lowin

Introduction

By Nicholas J. Rasmussen and Johannah Lowin

Anders Breivik's 1,518-page document released prior to the killing of 77 people at a government center and summer camp in Oslo, Norway in 2011. Brenton Tarrant's essay published before the mass shooting of 55 people at two mosques during Friday prayer in Christchurch, New Zealand in 2019. Patrick Crusius's tract posted before the massacre of 23 people at the Walmart in El Paso, Texas later that year.

All three of these detailed online manifestos are widely recognized as terrorist and violent extremist propaganda and disseminated among online sympathizers and members of white supremacy groups. Yet none of these writings — though they incite and inspire violence[1] — currently qualifies for inclusion in the Global Internet Forum to Counter Terrorism's hash-sharing database of terrorist and violent extremist content.

The reason for this is twofold: First, when several leading tech companies came together five years ago to establish the hash-sharing database, they agreed to include hashes of only a narrow subset of content. To find common ground, the original scope of the database was limited to material associated with organizations on the United Nations Security Council's Consolidated Sanctions List. In practice, that meant that nearly all hashes reflected content related to al-Qaeda, the Taliban, the Islamic State, or other groups that the United Nations had designated as terrorist organizations. As writings penned and published by far-right extremist attackers, none of the above manifestos met those criteria.

Second, the database in its current form only includes hashes of images or video and not PDF documents, which is how manifestos tend to surface and circulate online. Although the database has evolved over time to include hashes of content related to three specific attacks that have triggered GIFCT's Content Incident Protocol — a set of procedures developed to hasten the removal of content from a live-streamed event — that does not include PDFs.[2] As a result, while hashes of perpetrator-produced video footage of the devastating Christchurch attacks can be shared among members in the database, hashes of the perpetrator's manifesto — which the El Paso shooter cited as direct inspiration several months later — cannot.

[1] Like many such manifestos, the three mentioned here go so far as to feature detailed instructions for others to follow and/or careful reasoning for specific tactical choices.

[2] In order for a CIP to be activated, all four of the following conditions must be met: 1) a real-world terrorist, violent extremist, or mass violence event; 2) has been recorded or broadcast via livestream; 3) depicting murder or attempted murder; and 4) is being distributed on GIFCT member platforms or so broadly online that such distribution appears inevitable. Since the attacks in Christchurch, New Zealand in March 2019 initiated the creation of the CIP process, GIFCT has activated the CIP twice in response to offline attacks when perpetrator-produced content was shared on GIFCT member platforms: in Halle, Germany in October 2019 and in Glendale, Arizona in May 2020.

Unfortunately, the lifecycle of a terrorist manifesto rarely ends with the attack. Instead, these writings become like "the baton in a relay race of extremists, passed from one terrorist murderer to the next through online communities."[3] For instance, Tarrant's manifesto cited Breivik's attack as an inspiration and also claimed knowledge of manifestos by other far-right extremists, such as Dylann Roof, the perpetrator of the Charleston, South Carolina church shooting in 2015.[4] Following the Christchurch attack, Tarrant's manifesto was subsequently cited by Crusius, the El Paso shooter,[5] and an unnamed Singaporean teenage far-right extremist whose attack was foiled by authorities.[6] This pattern is common throughout the far-right extremist milieu as terrorist manifestos often cite, credit, and praise predecessors. Allowing terrorist manifestos to proliferate across the internet — unfettered and readily accessible — therefore contributes to the violent extremist culture of citation and heightens the risk of future attacks.

A Rights-Based, Multi-Stakeholder Approach:

Manifestos are but one example of the sort of harmful content linked to terrorism and violent extremism that GIFCT's hash-sharing database can and must evolve to address. One year ago, at the first Global Summit of the newly independent GIFCT, our team committed to leading a global multi-stakeholder effort to explore how to expand the hash-sharing database's utility, reach, and impact.

As part of this year-long initiative, GIFCT put human rights — including, but not limited to, freedom of speech and privacy — at the forefront of the process. At the advice of key civil society stakeholders involved in a range of GIFCT work and initiates, GIFCT engaged the firm Business for Social Responsibility (BSR) to undertake a forward-looking human rights impact assessment of the organization at this foundational moment. The recommendations in that report, released in full last week, helped shape and inform every element of this effort.

Chief among the report's findings was a widespread view among our stakeholder community that the narrow scope of the database reflects broader discrimination and bias in the counterterrorism field, specifically a disproportionate focus on Islamist extremist content rather than white supremacist content.[7] Indeed, the assessment identified GIFCT's review of the hash-sharing database's taxonomy as a key opportunity "to proactively address bias in the counterterrorism field."[8] As GIFCT increasingly aims to focus on behavior and content, in addition to dangerous individuals and organizations, this project seeks to do

[3] J.M. Berger, "The Strategy of Violent White Supremacy is Evolving," The Atlantic, August 7, 2019.
[4] Ari Ben Am and Gabriel Weinman, "Fabricated Martyrs: The Warrior-Saint Icons of Far-Right Terrorism," Perspectives on Terrorism 14, no. 5 (October 2020): 140.
[5] Jacob Ware, "Testament to Murder: The Violent Far-Right's Increasing Use of Terrorist Manifestos," ICCT, March 20, 2020.
[6] Max Walden, "Singaporean teenager arrested for allegedly planning Christchurch-inspired machete attack on mosques," ABC News, June 27, 2021, Link
[7] "Human Rights Assessment: Global Internet Forum to Counter Terrorism," Business for Social Responsibility, 2021.
[8] Business for Social Responsibility, "Human Rights Assessment," 26.

just that.

The Series

The first piece in this collection, **"Practical and Technical Considerations in Expanding the GIFCT Hash-Sharing Database,"** offers a practical lens through which to view the insights and recommendations from the independent academic papers that follow. Co-authored by GIFCT Director of Programming **Erin Saltman** and Director of Technology **Tom Thorley**, this chapter draws on a series of interviews with each GIFCT member to review the technical feasibility of various approaches to expansion. As our diverse array of members must ultimately be able to implement an expanded taxonomy in order for it to be valuable, this initial chapter outlines the critical practical and technical questions for GIFCT to consider throughout this process.

The next article, written by academics **Dan Byman and Chris Meserole**, builds on work they produced for a GIFCT research initiative three years ago to provide a framework and set of recommendations for developing a more explicit and comprehensive taxonomy of terrorist and violent extremist content. Noting that determining what content should be shared within the hash-sharing database remains an unresolved challenge, this piece argues for moving towards a taxonomy that is agnostic to ideology, globally applicable, and centered on content in addition to actors. The paper concludes by proposing concrete ways of developing a common taxonomy for content and actors while also addressing some of the potential implementation and process challenges ahead.

The recommendations that Byman and Meserole outline in their paper draw on a range of innovative ideas presented in a series of independent papers from international experts that GIFCT commissioned at the start of this year. Of the more than a dozen proposals submitted in response to a Request for Proposals on our website, GIFCT selected papers from academics, practitioners, and civil society leaders at the **University of Queensland and the Australian Muslim Advocacy Network,** the **Institute for Strategic Dialogue** (ISD Global) in London, the **University of Maryland's START program**, and **Hedayah** in Abu Dhabi, as well as a team of academics and practitioners from GIFCT's own Independent Advisory Committee. By bringing together authors with a wide array of methodological expertise, analytical approaches, and regional specialization, this series offers what Byman and Meserole describe as "the most original and comprehensive thinking to date on what criteria should be used for cross-platform hash-sharing."

Taking into account this diversity of perspectives and approaches, this volume concludes with a **series of initial next steps for GIFCT** to take in order to make the database more relevant and responsive to the global terrorist and violent extremist challenges we face today. Based on the ideas outlined in the research papers, feedback from our members, and our own technical analysis of what is feasible under the current architecture of the database, the conclusion elaborates on GIFCT's intent over the coming months to begin expanding the taxonomy of the database through the addition of three new categories — including attacker manifestos. Going forward, GIFCT will use the research and recommendations in this publication as a key resource for continuing to make iterative progress to expand the value of the hash-sharing database over time.

At the end of this volume, we also include an **appendix on select global definitions of terrorism and violent extremism** that we think are relevant to this ongoing work. As there is no universal agreement on the definition of terrorism, and far less on the parameters and legal definitions of violent extremism, we hope that this section will serve as a useful resource in bringing together definitions and approaches that GIFCT stakeholders have adopted. After providing some context on the etymology of the terms "terrorism" and "violent extremism," this appendix presents an overview of definitions used by governments represented on GIFCT's Independent Advisory Committee, as well as GIFCT member companies. We see this entire effort as a first step towards using GIFCT's convening power to create a "common understanding" of terrorism and violent extremism, a key recommendation in BSR's recent assessment of our work.

As we seek to broaden and diversify our membership, and as the nature of terrorist and violent extremist activity online evolves, GIFCT must focus on developing rights-based technical solutions that are thoughtful, practical, and scalable. In our view, hash-sharing is an innovative and important first step in GIFCT's efforts to share information among member companies and thereby promote more effective content moderation. But hash-sharing is only one piece of the puzzle. When it comes to technical solutions and industry collaboration to address terrorist and violent extremist exploitation of the internet, much more remains to be done. GIFCT looks forward to working with our diverse stakeholder community to advance this collaborative effort in service of a safer and more open internet and world.

Nicholas J. Rasmussen is the inaugural Executive Director of the Global Internet Forum to Counter Terrorism (GIFCT). Johannah Lowin is the Chief of Staff and Director of Strategic Initiatives at GIFCT.

GIFCT
Global Internet Forum
to Counter Terrorism

Background:

A Guide to the Taxonomy of GIFCT's Hash-Sharing Database

By GIFCT Staff

Background

A Guide to the Taxonomy of GIFCT's Hash-Sharing Database Pre-dating the establishment of GIFCT as an independent organization in 2020

In 2017, the founding members of GIFCT spearheaded a shared, safe, and secure industry database to house "perceptual hashes" of known terrorist-produced images and videos. Recognizing that there was not one agreed-upon international definition of terrorism, the original parameters of the hash-sharing database were limited to hashes of content (images and videos) that GIFCT members had removed from their services for being terrorist content and which were also produced by terrorist entities on the United Nations Security Council's Consolidated Sanctions List.[9]

To date, the only other hashes allowed in the hash-sharing database that do not correspond to entities on the U.N. list relate to content created by a perpetrator or accomplice of a terrorist incident when a Content Incident Protocol (CIP) is declared. The GIFCT CIP was developed in April 2019 in response to the tragic terrorist attacks in Christchurch, New Zealand. The CIP is a system that aims to thwart the online proliferation of content produced by a perpetrator during a real-world attack. As of July 2021, GIFCT has initiated the CIP twice since the New Zealand attacks in response to two separate real-world events.[10]

Hashes in the database are labeled per the following taxonomy:

UN List_Imminent Credible Threat (ICT): A public posting of a specific, imminent, and credible threat of violence toward non-combatants and/or civilian infrastructure.

UN List_Graphic Violence Against Defenseless People: The murder, execution, rape, torture, or infliction of serious bodily harm on defenseless people (e.g. prisoner exploitation, obvious non-combatants being targeted).

UN List_Glorification of Terrorist Acts (GTA): Content that glorifies, praises, condones, or celebrates attacks after the fact.

UN List_Recruitment and Instruction (R&I): Materials that seek to recruit followers, give them guidance, or instruct them operationally.

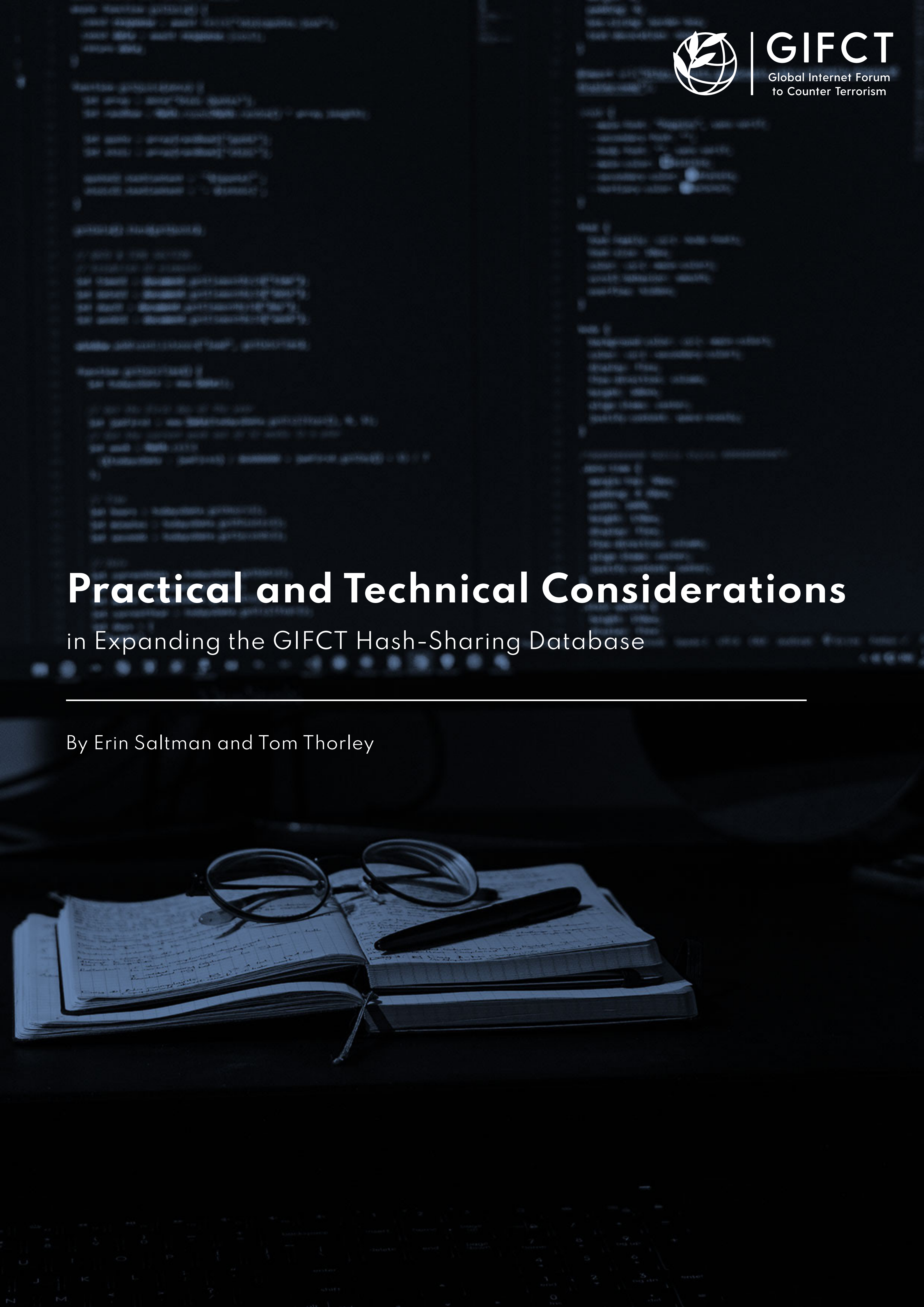
CIP_New Zealand Perpetrator Content: On March 15, 2019, GIFCT set a new precedent in the wake of the New Zealand terrorist attack. Due to the virality and cross-platform spread of the attacker's manifesto and attack video, and because New Zealand authorities deemed all manifesto and attack video content illegal,

[9] "United Nations Security Council Consolidated List," United Nations Security Council, accessed in June 2021, Link
[10] GIFCT commits to working collaboratively across industry, governments, and NGOs on protocols for responding to emerging or active events. More information on the CIP can be found here

GIFCT created a crisis bank in the hash database to help mitigate the spread of this content.

CIP_Halle, Germany, Perpetrator Content: On October 9, 2019, GIFCT activated the CIP following the shooting in Halle, Germany, and the perpetrator's attack video circulating on multiple digital platforms.

For more about the hash sharing database and how it operates please visit here.



GIFCT
Global Internet Forum
to Counter Terrorism

Practical and Technical Considerations

in Expanding the GIFCT Hash-Sharing Database

By Erin Saltman and Tom Thorley

Practical and Technical Considerations in Expanding the GIFCT Hash-Sharing Database

By Erin Saltman and Tom Thorley

Abstract

GIFCT prides itself on incorporating a multi-stakeholder approach in expanding and evolving efforts to counter terrorism and violent extremism online. The reports in this document highlight a range of important perspectives from international experts, practitioners, and civil society organizations. While this input is necessary to ensure approaches are lawful, legitimate, and proportional to the threat, these efforts at expanding and evolving GIFCT's response also need to work with the systems and operations of GIFCT member companies who ultimately benefit from a shared technical approach such as the hash-sharing database.

To ensure that GIFCT combines expert feedback with technically feasible approaches, in Feb 2021 GIFCT conducted one-on-one interviews with all GIFCT members. The interviews included a discussion of options and the range of approaches for expanding the hash-sharing database taxonomy, as well as potential future areas of work regarding technical approaches. This chapter focuses on the hash-sharing database taxonomy and approaches to its expansion with regard to policy challenges, technical challenges, adversarial shifts, and the viability of solutions.

The Hashing Process

When a hash-sharing database member identifies an image or video on their platform that has violated their terms of service and is associated with an U.N.-designated terrorist entity, they can produce a hash of the content and upload the hash to the hash-sharing database. These hashes are digital signatures of the image or video. Hashes are numerical representations of original content and cannot be reverse-engineered to recreate the image or video. The hashes used in the GIFCT hash-sharing database are "perceptual hashes" - which means visually similar content creates hashes that are mathematically close to each other.

The main algorithm used in this process is an open-source one developed and released by Facebook[11] called PDQ[12] (though the hash-sharing database also supports Microsoft's PhotoDNA).[13] The process (using an image such as the example below) first converts the picture to grayscale and resizes it so that all images are identically formatted before being hashed.

[11] Facebook, facebook/ThreatExchange. GitHub (November, 2020), link

[12] Facebook, facebook/ThreatExchange. GitHub (October, 2020), link

[13] Microsoft, PhotoDNA (2020), link

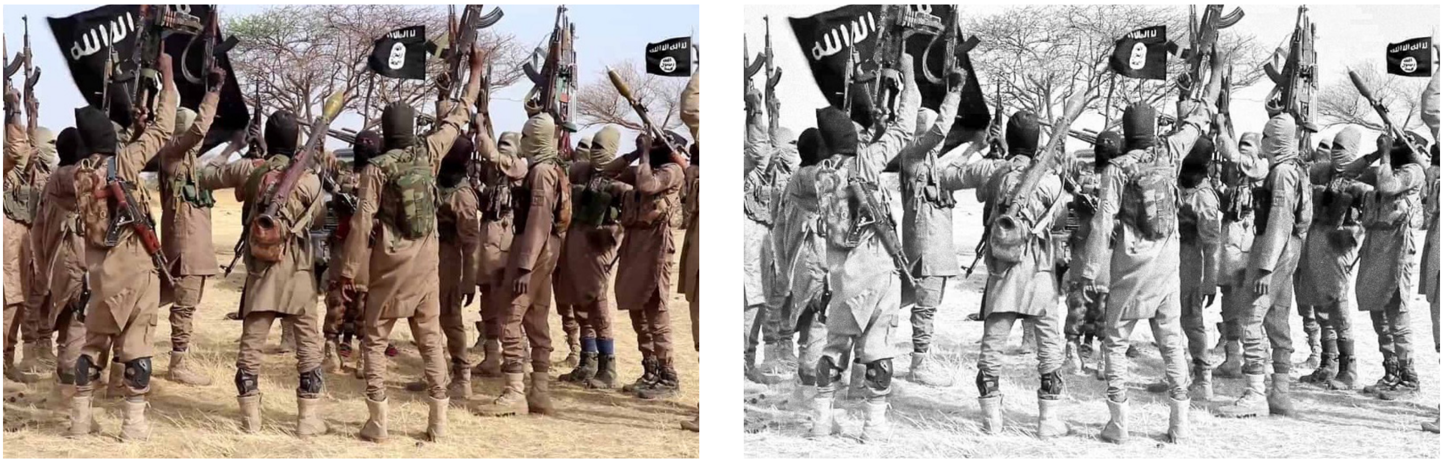


Fig 1. Image converted to grayscale and resized

To generate the hash a mathematical procedure known as Discrete Cosine Transform is applied to the small grayscale image, resulting in a 256 bit hash with hamming distance[14] and a quality metric that describes the level of detail in the image (featureless images have a quality of zero) such as the example below.

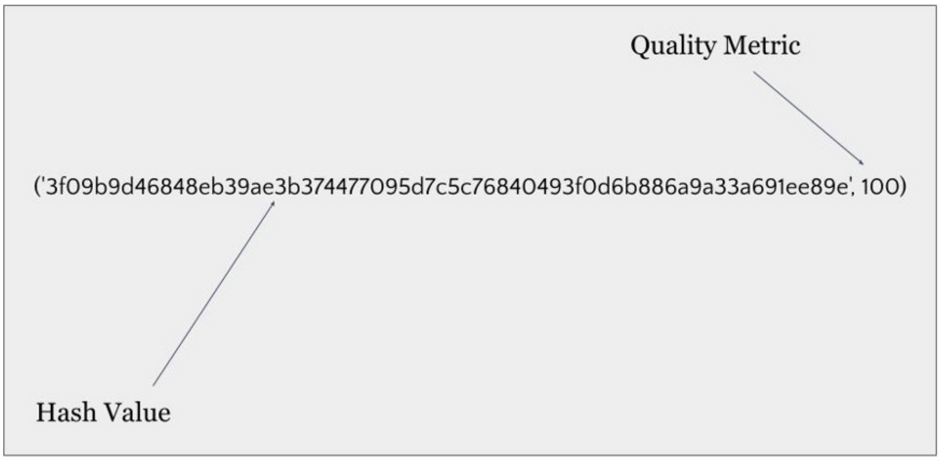

Quality Metric
('3f09b9d46848eb39ae3b374477095d7c5c76840493f0d6b886a9a33a691ee89e', 100)
Hash Value

Fig 2. Example PDQ hash generated from the image in Fig 1.

The hash-sharing database also supports hashing of video through algorithms such as TMK[15] (which builds on PDQ) and PhotoDNA.

Hashes allow GIFCT members to quickly identify visually similar content on their own platform which has been removed by one member, enabling them to review (or re-review) such content to see if it breaches their terms and conditions (without sharing any user data between companies).

When GIFCT members review the content that has been identified by matching it against hashes, they also have the option to give feedback to the system and tell other members whether they agree or disagree that any one hash relates to terrorist activity and rate its severity. GIFCT respects that each member has different policies, corporate purposes, and terms and conditions. As a result, there is not a one-size-fits-all approach to how companies use hashes to support their platforms or

[14] Pinch, R., "Hamming Distance – Encyclopedia of Mathematics," Encyclopedia of Mathematics (September 17, 2016), link
[15] Facebook, facebook/ThreatExchange. GitHub (August 27, 2020), link

how member companies apply their policies to the material surfaced from matches against hashes in the hash-sharing database.

As of July 2021, GIFCT's Transparency Report[16] shows that 320K distinct photos and videos have been hashed and shared through in the database. The following shows the breakdown of what kind of hashed content has been ingested into the shared database based on the existing database taxonomy.

Hashed content related to the U.N. designated entities list:

- Imminent Credible Threat: 0.1%
- Graphic Violence Against Defenseless People: 14.3%
- Glorification of Terrorist Acts: 77.2%
- Radicalization, Recruitment, Instruction: 1.7%

Hashed content related to perpetrator content associated with a Content Incident Protocol:

- Christchurch, New Zealand, attack and Christchurch, New Zealand, Perpetrator Content: 5.1%
- Halle, Germany, Perpetrator Content: 1.5%
- Glendale, Arizona, U.S., Perpetrator Content: 0.1%

Because GIFCT hosts a consortium of companies working together, GIFCT is not a social media platform itself and does not own or store any original source data or privacy data of any users associated with platform members.

Policy Challenges

GIFCT members represent an increasingly diverse set of platforms.[17] Members include companies large and small, U.S. and non-U.S. based, but also platforms that offer a range of different services. Some are more recognizable social media focused services while others cater to content storage, online marketplaces for goods or services, and private end-to-end encrypted communications platforms. GIFCT looks forward to the further inclusion of a variety of global tech companies and services that face terrorist and violent extremist exploitation. However, recognizing the heterogeneity of the internet brings up a range of policy challenges.

For any platform, the decision to keep up or remove content or accounts is based on a platform's unique terms of service and policy guidelines, often with reference to legal requirements in a given country where the company is based. These policies tend to be typically stated in a legal format in

[16] GIFCT, "GIFCT Transparency Report – July 2020," GIFCT Transparency Report (July 2020), link
[17] See the GIFCT Membership Criteria and Member Companies on the GIFCT Membership Page: link

the terms of service, but in layperson terms in the format of community standards or user guidelines.[18] These guidelines are put in place to ensure users can access the platforms freely and safely while understanding the principles behind when and why certain content, actions, or accounts might be restricted or removed.

These policies are generally global. Most companies design, review and update these policies based on feedback from a range of stakeholders that might include their users, government bodies, and global experts in fields such as technology, public safety, and human rights. GIFCT members enforce their own respective policies and practices in response to violations of their terms of service or standards. While there is no one globally agreed-upon definition of terrorism or violent extremism, most tech companies have independently developed definitions and approaches based on government and expert resources and in consideration of what will work best based on how their platform operates and what sort of signals they have access to in order to assess violations.

Definitions of Terrorism

Any discussion of how to expand the taxonomy of the hash-sharing database inevitably returns to the root question of how to define terrorism and violent extremism – specifically with reference to defining parameters for the type of content or signals emerging online. Just like governments and international institutions, there is not one agreed-upon definition of terrorism or violent extremism for private companies. Many companies defer to government designation lists to ensure compliance with legal obligations, while others have developed their own working definitions in order to assess groups based on behavioral signals and offline incidents.

To date, GIFCT allows for a broad discussion and interpretation of terrorism and violent extremism within its programmatic and research efforts, but a narrow definition for inclusion in the hash-sharing database, since hashes have the potential to lead to source content on a given platform with possible repercussions for the user who shared or stored the content. The database currently includes a combination of a list-based approach (hashes related to terrorist organizations on the U.N. list) and a behavior plus content-based approach (attacker or perpetrator content related to a Content Incident Protocol (CIP)).

In discussions with GIFCT member companies about ways that the taxonomy could be expanded, the debate centers on whether such an expansion should proceed with a list-based approach or a behavioral and content-focused approach. Lists inherently focus on organizations and individuals while behavior approaches focus on the specific type of content (that might appear from a terrorist or violent extremist organization). List-based approaches are generally positive with regard to transparency in openly indicating the groups implicated, but remain highly political and often geographically or ideologically biased. Behavior-based approaches offer a more nuanced path for recognizing adversarial shifts and other forms of violent extremism but can be difficult to maintain and clearly define.

[18] See the GIFCT Member Resource Guide to review all members' policies on terrorism and violent extremism: link

List-based Approaches

In discussions about expanding government list-based approaches, companies expressed both concerns about Western biases as well as concerns about potential pressure from non-democratic countries to use lists that might not comply with human rights frameworks. There are cases where countries have added activist networks, journalists, or opposition parties to "terrorist lists" as a means to suppress government opposition. There was also acknowledgment by companies and other stakeholders that these lists all tended to focus on Islamist extremist terrorism (despite some recent additions of groups related to white supremacy in the U.S., U.K., and Canada). The other issue for companies trying to work with list-based approaches was that it remains very difficult to unequivocally identify and label members of a terrorist or violent extremist group unless there are very obvious profile or account indicators or self-declared membership.

Behavioral-based Approaches

Companies pointed to a range of examples where seemingly clear violations might not be so easy to act against in the online space. For example, bomb-making instructions are also legal and normative in many online forums discussing hunting, military training, or even firework making. Instructions on how to cause physical harm or death to another human are commonplace in self-defense, activist, and military forums. Even if some companies could agree that this type of content might be removed under other policies, in both examples there is the potential for scope creep that goes above and beyond the GIFCT parameters of focusing on terrorism and violent extremism.

Thus, GIFCT aims to focus on developing categories for inclusion in the hash-sharing database that are (1) specific to terrorism and violent extremism, (2) recognize how content specific to these actors and groups manifests and spreads online, and (3) involve processes that can lean into expert scrutiny and transparency. While GIFCT currently does not have the in-house capacity to build, own, and maintain a global list of terrorist and violent extremist organizations and perpetrators, this is something GIFCT as an independent NGO could work towards, so long as the necessary transparency and clear definitional parameters were in place. For now, this means expansion of the hash-sharing database should focus on clear behavioral and content-focused labels or lean into existing lists and structures.

Tech Challenges

GIFCT's initial focus was helping companies moderate user-generated content (UGC).[19] While this is still a core part of GIFCT's focus, as its membership has grown (from the original four members there are now seventeen) the forms and functions of its member companies have diversified. Though they agree on the core GIFCT membership pillars,[20] such as a commitment to fundamental human rights and annual transparency reporting, each member company has a different corporate purpose and philosophy. Some members are end-to-end encrypted (E2EE)[21] platforms while other member

[19] Krumm, J., N. Davies, and C. Narayanaswami, "User-Generated Content," IEEE Journals and Magazine | IEEE Xplore (October, 2008), link
[20] GIFCT, Membership (February 10, 2021), link

companies operate in different parts of the technology stack (providing hosting or cloud services rather than user-facing functions). Some members solely deliver social media platforms while others deliver an array of services such as retail, financial services, gaming, web hosting, and more. This diversity helps extend GIFCT's reach and the impact that it can have as an organization devoted to fighting terrorism, but it is essential that GIFCT understands the complexity and differences in its membership and avoids one-size-fits-all solutions.

For companies that have platforms that support UGC, sharing hashes of images and videos that have been identified as being related to terrorist activity is one important step in preventing terrorists and violent extremists from exploiting their platforms (while protecting user privacy and being computationally efficient). Companies that are E2EE have no access to user communication content and so hashed images do not help them to identify terrorist activity on their platforms.

Despite this diverse array of companies, during interviews with member companies several challenges came up again and again. Laying out these challenges helps explain where the hash-sharing database fits in as a solution and how expanding it will help, but also what other significant gaps members highlighted and what viable solutions could potentially be explored. Also briefly considered are the key limitations and considerations that need to be addressed hand in hand with these potential solutions.

Protected Groups

GIFCT members recognize the need to protect artistic works, journalism, and academic work. Logos of terrorist groups could be part of a slide in a professor's slide deck or they could be attached to propaganda, so the context in which these logos appear is critical. While social media platforms tend to be able to check the user accounts associated with content that has been identified as terrorist-related, very few member companies have a real-name requirement for creating an account and many member platforms do not have any wider context about their users to draw from. As journalists and academics are routinely threatened and otherwise abused online by members of the terrorist groups that they seek to research, a level of anonymity for these users is critical to be able to carry out work to understand and counter violent extremist groups.

Diverse Languages

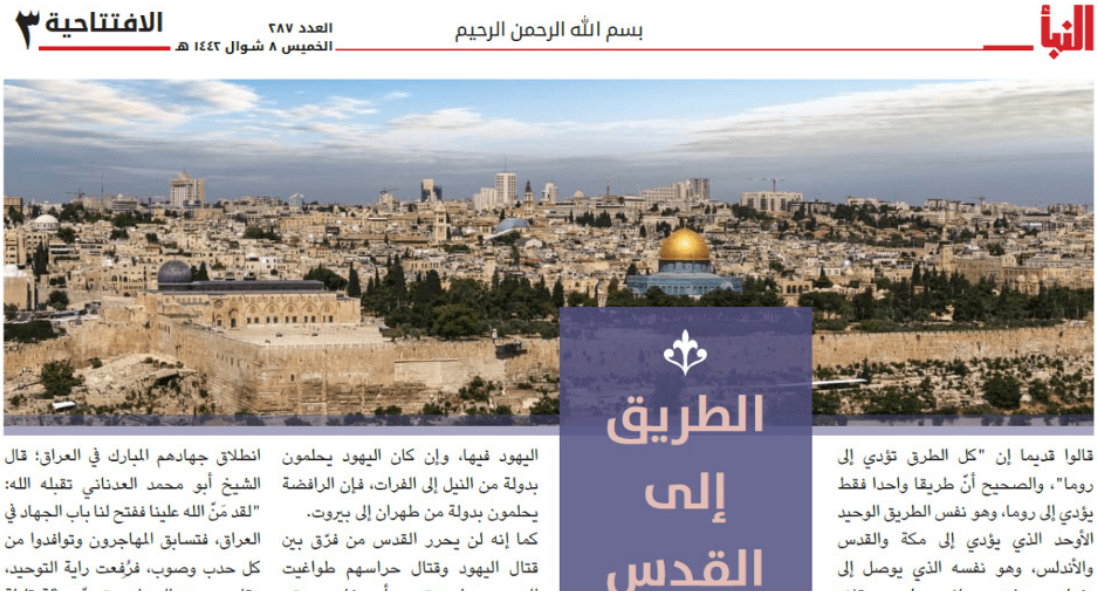
If who is posting presents a challenge to GIFCT members' companies, the diversity of the material posted provides a further challenge. Terrorist content is shared in an array of different languages; for example, "Al-Naba's infographics are systematically translated into English, French, Italian, Russian and other languages a few hours after the release in Arabic."[22] Although a few large companies have specialist teams with subject matter expertise and a wide range of language skills at their disposal, most companies have small teams to review content and very few linguists with the appropriate mix of rare dialects.

[21] "End-to-end encryption is a system of communication where the only people who can read the messages are the people communicating. No eavesdropper can access the cryptographic keys needed to decrypt the conversation—not even a company that runs the messaging service." Greenberg, A., "Hacker Lexicon: What Is End-to-End Encryption?," Wired (November 25, 2014), link
[22] Gluck, R., and L. Binder, "Trends in Islamic State's Online Propaganda: Shorter Longevity, Wider Dissemination of Content," ICCT (December 5, 2018), link

Diverse Formats: Images, Videos, Text, and Audio

Terrorist groups also use a wide variety of media formats. In particular, audio has been a mainstay of violent extremists for many years, using radio to reach local audiences[23] and releasing audio clips online (for example the 2018 ISIS release of an audio message purportedly of its leader Abu Bakr al-Baghdadi).[24] With the rise of social media networks based on voice,[25] use of this medium for terrorist purposes is only going to grow. The current hash-sharing database does not include sharing of information related to audio and the technical tools available to companies across the sector vary dramatically. Although advanced open-source speech-to-text and audio classification systems are available,[26] integrating these and having the appropriate training data available remains challenging (especially for smaller companies). Audio is also very time consuming for teams that review content to analyze and requires a high degree of subject matter expertise and contextual knowledge (for example understanding the nasheed compositions used in ISIS propaganda videos).[27]

Although technically simpler to deal with than audio, text is also used extensively by terrorists and can pose a challenge for companies to deal with. Publications such as Al-Naba, ISIS's weekly newsletter, may contain images but are largely text-based.



النبأ

بسم الله الرحمن الرحيم

العدد ٢٨٧
الخميس ٨ شوال ١٤٤٢ هـ

الافتتاحية ٣

الطريق
إلى
القدس

قالوا قديما إن "كل الطرق تؤدي إلى
روما"، والصحيح أنّ طريقا واحدا فقط
يؤدي إلى روما، وهو نفس الطريق الوحيد
الأوحد الذي يؤدي إلى مكة والقدس
والأندلس، وهو نفسه الذي يوصل إلى

اليهود فيها، وإن كان اليهود يحلمون
بدولة من النيل إلى الفرات، فإن الرافضة
يحلمون بدولة من طهران إلى بيروت.
كما إنه لن يحرر القدس من فرّق بين
قتال اليهود وقتال حراسهم طواغيت

انطلاق جهادهم المبارك في العراق؛ قال
الشيخ أبو محمد العدناني تقبله الله:
"لقد مَنّ الله علينا ففتح لنا باب الجهاد في
العراق، فتسابق المهاجرون وتوافدوا من
كل حدب وصوب، فرُفعت راية التوحيد،

Fig.3 Al-Naba 287 page 3

While Natural Language Processing and other text analysis tools can be used to identify violent extremist content, the way in which text is used is highly context dependent. Many approaches to analyzing and classifying it can lead to high rates of false positives, especially as the length of the text decreases.[28] Text is also a challenge as most of the efforts of the tech industry to deal with

[23] "IS radio beams propaganda, threats across rural Afghanistan," AP News (January 21, 2016), link
[24] "Islamic State releases new audio, purportedly of its leader," AP News (August 23, 2018), link
[25] Basu, T., "The future of social networks might be audio," MIT Technology Review (January 28, 2021), link
[26] Google, google-research/leaf-audio, GitHub (March 5, 2021), link
[27] Alvi, H., "Musical Criminology: A Comparative Analysis of Jihadist Nasheeds and Narco Corridos," Air University (June 2020), link

online harms have been focused on image and video. (This is true not just in countering terrorism, but also in efforts to combat child sexual abuse online.) As a result, member companies have a wide variety of approaches to text, with some not scanning it at all, some manually checking text, some using text-string searching and some using machine learning to surface potentially violative text.

Live Streaming

One further challenge that GIFCT members are dealing with is the potential for terrorist use of live-streaming services. In response to this challenge and the tragic Christchurch terrorist attack in New Zealand in March of 2019, GIFCT's CIP was created to alert member companies of potential video content circulating online from a real-time terrorism or violent extremist event such that they could quickly assess and act on this information.[29]

By declaring a CIP, hashes of an attacker's video and other related content produced as part of the attack are shared in the GIFCT hash database with other GIFCT member platforms. Furthermore, a continuous stream of communication is established among all GIFCT founding members to identify and address risks and needs during an active CIP.

The CIP is one part of a response to terrorist exploitation of live-streaming video systems and hashing alone cannot address this challenge, as hashes are static snapshots and a livestream is by definition a dynamic and evolving piece of content. While the technology for analyzing live-streaming video is developing rapidly (and in some cases is available open-source) and companies have put in place many safeguards to mitigate the risk of abuse, there is still more to be done to address this challenge.[30]

Adversarial Shifts and GIFCT Member Considerations

GIFCT and its partners have a unique mission in approaching counter-terrorism and counter-extremism efforts from the perspective of tech companies as the primary stakeholder. In these efforts, it is crucial to understand what the adversarial shifts look like online in order to develop strategies that can adequately mitigate harm and prevent further exploitation. Expansion of the hash-sharing database taxonomy is one solution under consideration among a wider set of GIFCT programmatic and technical workstreams that are exploring how to address these changes. In cultivating viable solutions there are a few trends in adversarial shifts that GIFCT has considered.

Big Versus Small Companies

A range of academic publications and reports have highlighted that the terrorist and violent extremist threat is both transnational and cross-platform.[31] Bad actors, just like average users,

[28] Alrhmoun, A., S. Maher, and C. Winter, "Decoding Hate: Using Experimental Text Analysis to Classify Terrorist Content," GNet Research (September 2020), link
[29] GIFCT, Crisis Response (March 23, 2021), link
[30] Redmon, J., and A. Farhadi, "YOLO Live," Machine Learning for Artists (December 2016), link

utilize a multitude of platforms to get things done quickly and cheaply and choose platforms based on whether or not more private or public communication is preferable for various goals. GIFCT was founded by tech companies who recognized this shift, but it is also a reminder that any tools that are developed or taxonomy expansion needs to consider the capacities and manpower available to different types of companies in operationalizing these efforts. GIFCT members include some of the largest and smallest platforms with vastly different resources both in terms of tooling and human resource.

As larger companies put more effort towards counter-terrorism and countering violent extremism, these increasingly proactive efforts lead to two shifts; bad actors hiding their intentions on the larger platforms to evade detection and removal, and migration to smaller, less regulated platforms where they can be more overt without suffering negative consequences. All GIFCT efforts need to ensure that they are catering to different capacities. This includes providing tangential support to companies' tech, policy and operational teams where necessary, so that tooling and taxonomies are both understood and can be implemented in line with a company or platform's existing policies and procedures.

In-House Subject Matter Expertise

While a company can be assessed in part by its size and capital, even some of the largest companies are new to confronting terrorist or violent extremist exploitation on their platforms. There are many companies that do not have in-house expertise on terrorism or violent extremism, making it difficult to develop more nuanced moderation practices. While some of the larger companies have hired in-house teams, others rely on third-party expertise or default to moderation guidelines that tackle only the most obvious content or signals. It is also worth remembering that companies have to build policy and practices that cover a much wider range of harms above and beyond terrorism and violent extremism. Often harm types like spam, copyright, or hate speech will be much higher in volume. Nevertheless, terrorism and violent extremism will continue to be a low-prevalence but high-risk area of concern.

As such, GIFCT will continue to think through ways of pairing its technical advancements with wider knowledge-sharing efforts and provide access to resources that can facilitate moderation teams' understanding with context. Insights and contextual knowledge will continue to be of great importance as tech companies expand to have a transnational global reach and increasingly adopt policies and practices that reflect the nuances of how terrorism and violent extremism manifests in different parts of the world. This is why the work of the Global Network on Extremism and Technology (GNET), with its constant insight briefings, as well as the Knowledge Sharing Platform run by Tech Against Terrorism, are important to support. These tools and resources give platforms the action-oriented insights and context to apply GIFCT tools.

[31] See Fisher, A., N. Prucha, and E. Winterbotham, "Mapping the Jihadist Information Ecosystem," Global Research Network on Terrorism and Technology, 6 (2019), link ; Clifford, B., "Migration Moments: Extremist Adoption of Text-Based Instant Messaging Applications," Global Network on Extremism and Technology (November, 2020), link

Companies Willing to Work Collaboratively Versus Isolationist Platforms

National and international counter-terrorism and counter-extremism efforts and research have highlighted the wide range of platforms being exploited. However, as multi-stakeholder efforts have formed to combat this digital exploitation, there is a clear divide between companies that are willing to work collaboratively to problem-solve, and those that stay away from joint efforts – and some openly advocating against any self-regulation of speech. Often, the wider public focus remains on the few larger companies that do come to the table, despite the known proliferation of abuse to a myriad of smaller or lesser-known platforms.[32]

As larger and more resourced companies increase their proactive efforts, this proliferation to smaller platforms will continue. While GIFCT and other collaborative approaches to tackle harms have an open door to new members – and also provide a range of open-source resources – there are increasingly aware of spaces online that these approaches will not be able to reach. What to do with companies unwilling to engage in dialogue and solution-building is a question for governments and practitioners that has yet to be fully answered.

Viability of Solutions

Given the challenges listed above, the need to expand the taxonomy of the hash-sharing database and the variety of formats that can be shared within it – while improving transparency and accountability – is essential. However, even the most robust hash-sharing system is only one part of what is required to counter terrorist and violent extremist activity online.

Context and Domain Knowledge

One theme that consistently came up during our interviews with members was providing context and domain knowledge. The teams that review material once it has been flagged due to being a hash match, or for another reason such as being reported by a user, vary dramatically in the size, knowledge, and skills that they have available. Even for the most experienced and well-resourced teams, the pace of adversarial shifts is a challenge. Resources that can help teams understand the context of a piece of potentially violative material are crucial, such as Tech Against Terrorism's Knowledge Sharing Platform.[33] Keeping these resources up to date and available is a critical pillar in the effort to fight terrorist exploitation of online platforms and also helps teams improve their decisions, leading to fewer false positives and less over-removal of content.

Logos and Symbols

The use of logos and symbols by various groups, especially given the adversarial shift and their sheer number, presents a unique challenge for content moderation teams. The guide to online radical-right symbols, slogans, and slurs[34] by the Centre for Analysis of the Radical Right (CARR) highlights the complexity in the use of such symbols by the extreme right, and there is a similar usage

[32] Clifford, "Migration Moments."
[33] Tech Against Terrorism, "Knowledge Sharing Platform," Knowledge Sharing Platform (2021), link
[34] Richardson, J., "CARR Guide to Online Radical-Right Symbols, Slogans and Slurs – Centre for Analysis of the Radical Right," Centre for Analysis of the Radical Right (May 4, 2020), link

of logos by other terrorist groups. These logos are used in content shared by extremists but also in the profile picture of accounts used by extremists online. In addition to the fact that these logos are related to terrorist groups and the context around them, more robust efforts can be developed to identify, extract and share signatures associated with these logos between companies to help with the detection of terrorist activity.

Machine Learning and Artificial Intelligence

Machine learning and artificial intelligence are often talked about as a panacea for all ills. Within many member companies, machine learning systems are used very successfully for identification and classification of terrorist and violent extremist content. While this especially helps larger companies who have the resources and training data to implement such systems, this is not necessarily the case for smaller companies and does not help with collaboration across platforms. While there are potential theoretical approaches to using machine learning across different organizations (such as federated learning), much work is needed to explore the legal, ethical, and human rights implications of such systems before they should be put into operation.

Practical steps

The Conclusion and Initial Next Steps chapter culminating this series of papers uses the findings from tech companies outlined in this chapter as well as the wider discussion papers to describe the incremental next steps in the expansion of the hash-sharing database. The expansion of hashed categories to include (1) attacker manifestos, (2) PDFs of branded terrorist content, and (3) TCAP URLs shows a crucial (though limited) first step in broadening scope. Some of these three new categories help break GIFCT out of inherent list-based biases, as they recognize how terrorist and violent extremist content manifests and is shared across platforms, and builds out new capacities in the form of content that can be hashed and shared. That being said, this change remains limited so that the new categories can be easily defined, explained, and scaled.

Beyond expanding the taxonomy of the hash-sharing database, the next steps for GIFCT's technical efforts should be focused on addressing the challenges of member companies outlined above and ensuring improved transparency. In the short term, developing approaches to hashing text such as TLSH[35] and use of tools such as Tesseract (OCR)[36] will enable hashing of manifestos linked to terrorist attackers, which member companies all agree violate their existing terms and conditions. In the medium term, approaches to hashing of audio, feature extraction from images, and logo detection should be considered. Finally, in the long term, efforts should be made to explore the art of the possible with regard to machine learning with a focus on how to mitigate bias and address ethical and human rights concerns.

[35] Oliver, J., "TLSH – A Locality Sensitive Hash," Tlsh.Org (March 13, 2021), link
[36] Google, Projects, Opensource, Google (December 26, 2019), link

Bibliography

Alrhmoun, A., S. Maher, and C. Winter. "Decoding Hate: Using Experimental Text Analysis to Classify Terrorist Content." GNet Research (September 2020). link

Alvi, H. "Musical Criminology: A Comparative Analysis of Jihadist Nasheeds and Narco Corridos." Air University (June 2020). link

Basu, T. "The future of social networks might be audio." MIT Technology Review (January 28, 2021). link

Clifford, B. "Migration Moments: Extremist Adoption of Text-Based Instant Messaging Applications." Global Network on Extremism and Technology (November, 2020). link

Facebook. facebook/ThreatExchange. GitHub (August 27, 2020). link

Facebook. facebook/ThreatExchange. GitHub (October, 2020). link

Facebook. facebook/ThreatExchange. GitHub (November, 2020). link

Fisher, A., N. Prucha, and E. Winterbotham. "Mapping the Jihadist Information Ecosystem." Global Research Network on Terrorism and Technology, 6 (2019). link

GIFCT. "GIFCT Transparency Report – July 2020." GIFCT Transparency Report (July 2020). link

GIFCT. Membership (February 10, 2021). link

GIFCT. Crisis Response (March 23, 2021). link

Gluck, R., and L. Binder. "Trends in Islamic State's Online Propaganda: Shorter Longevity, Wider Dissemination of Content." ICCT (December 5, 2018), link

Google. Projects. Opensource. Google (December 26, 2019). link

Google. google-research/leaf-audio. GitHub (March 5, 2021). link

Greenberg, A. "Hacker Lexicon: What Is End-to-End Encryption?" Wired (November 25, 2014), link

"IS radio beams propaganda, threats across rural Afghanistan." AP News (January 21, 2016), link

"Islamic State releases new audio, purportedly of its leader." AP News (August 23, 2018), link

Krumm, J., N. Davies, and C. Narayanaswami. "User-Generated Content." IEEE Journals and Magazine | IEEE Xplore (October, 2008), link

Microsoft. PhotoDNA (2020), link

Oliver, J. "TLSH – A Locality Sensitive Hash." Tlsh.Org (March 13, 2021), link

Pinch, R. "Hamming Distance – Encyclopedia of Mathematics." Encyclopedia of Mathematics (September 17, 2016), link

Redmon, J., and A. Farhadi. "YOLO Live." Machine Learning for Artists (December 2016), link

Richardson, J. "CARR Guide to Online Radical-Right Symbols, Slogans and Slurs – Centre for Analysis of the Radical Right." Centre for Analysis of the Radical Right (May 4, 2020), link

Tech Against Terrorism. "Knowledge Sharing Platform." Knowledge Sharing Platform (2021), link

United Nations. United Nations Security Council Consolidated List. United Nations Security Council (2021), link

GIFCT
Global Internet Forum
to Counter Terrorism

Expanding the Hash-Sharing Database

By Daniel Byman and Chris Meserole

Expanding the Hash-Sharing Database

By Daniel Byman and Chris Meserole

In December of 2014, an F-16 piloted by Lt. Moath al-Kasasbeh of Jordan crashed into the Syrian countryside outside Raqqa and he was captured by the Islamic State.[37] Kasasbeh's fate remained unknown until the Islamic State issued a demand that Jordan exchange Kasasbeh for Sajida al-Rishawi, whom Jordan sentenced to death for attempted terrorism. When Jordan refused, the Islamic State executed Kasasbeh with trademark cruelty, burning him alive in a cage. Yet the Islamic State didn't just kill Kasasbeh: the group also filmed the execution and published the video on the internet. A link to the video was released on Twitter and spread to other platforms.[38] Within hours, some major platforms scrambled to remove the video from their services while others kept it up or allowed stills from the video in the name of free expression. For many days after the incident, the video and other Islamic State content remained widely available online.[39]

At the time, the struggle to remove the video illustrated the difficulty of real-time cross-platform information sharing about terrorist and violent extremist content (TVEC). The video violated the terms of service of many major social media platforms and file-sharing services, but some emphasized free expression or had different procedures that complicated cooperation among them. Finding and removing the content – and doing so again and again for each slight variant of the video that emerged – was a labor-intensive and complicated process. Large platforms found themselves playing whack-a-mole against each version of the video that surfaced. This was difficult enough for well-resourced platforms like YouTube, but it proved nearly impossible for smaller platforms, which typically lacked the expertise and resources necessary to locate the video on their own. These platforms found themselves reliant almost entirely on their users to report the video. Compounding the problem was that the Kasasbeh video was just one piece of content amid a veritable deluge of Islamic State propaganda: the group rose to prominence in part because it flooded major social media platforms with more content than they had the capacity to remove.[40]

As the Kasasbeh and other videos have shown, the challenge of efficiently removing TVEC on any one platform is deeply entwined with the challenge of doing so across platforms. The more efficiently platforms can share TVEC material, the less exposure that material will have both on any one platform and across the industry overall. Cross-platform sharing thus benefits each individual platform as well as the broader sector of social media networks

[37] Laura Smith-Spark and Michael Martinez, "Who was Jordanian pilot Moath al-Kasasbeh, killed by ISIS?" CNN, February 3, 2015, link

[38] "Jordan pilot hostage Moaz al-Kasasbeh 'burned alive'," BBC News, February 3, 2015, link

[39] Jane McCallion, "Facebook refuses to remove images of Jordanian pilot murder," CRN, February 5, 2015, link; Counter Extremism Project, "Violent Extremist Content Still Found on YouTube," July 26, 2017, link

[40] J.D. Maddox, "Lessons from the Information War: Applying Effective Technological Solutions to the Problems of Online Disinformation and Propaganda," George Washington University Program on Extremism (September 2019), link

and file-sharing services.

Yet if the benefits of cross-platform sharing are clear, the two main impediments to doing so may be less obvious. The first is that platforms are restricted from sharing user content with unauthorized third parties. YouTube, for example, cannot share a video with Facebook without violating user expectations of privacy, much less legal frameworks like the European Union's General Data Protection Regulation (GDPR).[41] Therefore, the only way cross-platform sharing can work is for services to share anonymized unique identifiers of content rather than content itself. The other major challenge to cross-platform sharing is establishing a common definition of TVEC and a shared process for evaluating whether content meets that definition. There would be little value in sharing at all without a common understanding of what content should be shared, since platforms would stop relying on third-party information if they do not trust that it meets their criteria for moderating content.

The first of those challenges has a straightforward solution. Although platforms cannot share content with one another directly, they can share what are known as "hashes" – anonymized digital signatures of content. So long as each platform uses the same hashing algorithm for a given image or video, then they will all produce the same digital signature for that piece of content. And since the signature – typically a string of seemingly random numbers and letters – cannot be reverse-engineered to produce the original content, platforms can share them without sharing the underlying content itself. Given the advantages of hash-sharing, in 2016 several leading internet companies established a hash-sharing database for TVEC material, which was then placed under the management of the Global Internet Forum to Counter Terrorism (GIFCT) in 2017.

Yet if the reliance on hashes solves the first major challenge of cross-platform information sharing, it only heightens the importance of the second. Since platforms are only sharing signatures of content rather than the content itself, they need to trust that the signatures are generated from content that meets a common set of criteria. In other words, the platforms need to trust that they are all using the same definition and taxonomy for what TVEC material to share.

However, defining the criteria for what counts as TVEC is easier said than done. In contrast to other hash-sharing databases for harmful content, such as one maintained for Child Sexual Abuse Material (CSAM), hash-sharing for TVEC material is inherently more challenging. Just as "one man's freedom fighter is another man's terrorist," so too the same image or video may be viewed as terrorist content by one person and the legitimate political speech of a dissident or partisan by another. What content should qualify as "terrorist" is far more widely contested relative to most other online harms.

[41] Regulations like GDPR stipulate that companies cannot share data with third parties without user consent without removing all personally identifiable information and source content. However, companies are able to share user-generated content with officials and law enforcement following formal legal requests by governments.

To date, GIFCT's Hash-Sharing Database (HSDB) has addressed the challenge by including only a narrow subset of TVEC material. Indeed, as GIFCT explained in its latest Transparency Report, "To find common ground, the original scope of the hash-sharing database was … limited to content related to organizations on the United Nations Security Council's Consolidated Sanctions List."[42] As a result, nearly all the hashes currently in the database refer to content linked to the Islamic State, Al-Qaeda, the Taliban, and other groups designated as terrorist organizations by the United Nations. The only hashes not related to content affiliated with those groups come from the three attacks that have triggered the Content Incident Protocol (CIP), a set of procedures GIFCT developed to facilitate the removal of content from a livestreamed or recorded terrorist attack.[43]

Given concerns about free speech and expression, the HSDB's impulse toward restraint – and to including only content tied to the U.N. list or the CIP – seems both reasonable and appropriate. However, its reliance on the U.N. list creates at least two problems. The first is that (as the authors have pointed out in a prior paper) both the U.N. list and national terrorist designation lists are not the result of independent and objective processes but instead reflect political priorities.[44] As a result, they are overwhelmingly skewed toward terrorist groups that self-identify as Islamist, and include few groups with "far-right," ethnonationalist, or other non-Islamist ideologies. The second is that even though the HSDB is focused on terrorist and violent extremist content, the criteria for inclusion on such lists is actually centered on terrorist and violent extremist actors. Since many terrorist and violent extremist movements have shifted toward leaderless or loosely networked structures, the mismatch between content and actors means that the database will not include any TVEC produced by undesignated individuals and movements that lack a clearly defined organizational structure.

Determining what content should be shared within the HSDB thus remains an unresolved challenge.[45] Five years after it was first developed, the HSDB needs a more explicit and comprehensive taxonomy of the kinds of TVEC that should be shared within it. Ideally, that taxonomy should be agnostic to ideology, globally applicable, and centered on content in addition to terrorist or extremist actors. As a first step toward developing that taxonomy, we worked with GIFCT to commission a series of research briefs, which we briefly describe below. We then sketch a framework and set of recommendations of our own, recognizing that changes will require an involved and time-consuming process.

[42] GIFCT, "GIFCT Transparency Report July 2020," (July 2020): 2, link
[43] GIFCT, "Content Incident Protocol," (n.d.), link
[44] Chris Meserole and Daniel Byman, "Terrorist Definitions and Designations Lists: What Technology Companies Need to Know," Royal United Services Institute for Defence and Security Studies, (July 2019), link
[45] How hashes should be entered into the HSDB, as well as the broader processes by which the dataset should be managed, also remains an unresolved challenge. However, those process questions largely remain beyond the scope of this paper.

Overview of Other Papers in This Series

Although the HSDB's taxonomy initially focused solely on U.N.-designated groups, the problems with such an approach quickly became apparent. By relying on a list that only included violent Islamist groups, the HSDB would never be able to share hashes of terrorist groups with different ideologies such as white-supremacist groups. And by focusing solely on actors rather than content, the HSDB would also never be able to share hashes of standalone content unaffiliated with any group at all, even if it promoted terrorist or extremist violence. After the Christchurch attack in 2019, GIFCT took an early step at addressing both of those problems by developing the CIP for recorded or livestreamed attacks. Since the CIP can be triggered without reference to a known terrorist group, the hashed content it produces is not dependent on any actor affiliation, much less actors with a single ideology. Indeed, of the three incidents that triggered the CIP, none have included perpetrators with a violent Islamist ideology. In addition, the CIP helps GIFCT use the HSDB to respond to incidents closer to real time.

Yet even with the CIP, the fundamental challenges with the HSDB's existing taxonomy remain. To generate innovative ideas for how to improve on that taxonomy, GIFCT commissioned five papers that either propose a new taxonomy of their own or offer original insights on how such a taxonomy might be constructed. By drawing on authors with a wide array of methodological expertise, analytical approaches, and regional specializations, the series as a whole offers the most original and comprehensive thinking to date on what criteria should be used for cross-platform hash-sharing.

The first paper offers a new actor taxonomy of its own. Experts at the University of Queensland and the Australian Muslim Advocacy Network propose a "Dynamic Matrix of Extremisms and Terrorism" (DMET) as a way of encompassing a far wider range of extremist groups, both violent and non-violent, for technology companies to consider. DMET conceptualizes a continuum of extremism, ranging from simple and common partisanship to violent behavior and terrorism: a continuum that can be applied not just to existing groups in the HSDB like Al-Qaeda and the Islamic State but also to right-wing, separatist, and other causes, including behavior by state leaders. To determine where groups and networks are on the continuum, the authors identify various cognitive cues, organizational dynamics, and other markers such as the presence of dehumanization rhetoric and behavior that can be tracked and coded. Determinations would change as new information comes in or should the group change its behavior. The response of GIFCT member companies could vary based on where groups and individuals are on the spectrum and according to the differing standards of the companies. In addition, transparency would be necessary to ensure the integrity of the process and local cultural and social awareness is vital to reflect differing area standards.

By contrast, the Institute for Strategic Dialogue (ISD) offers a content-focused taxonomy that addresses one of the hardest problems facing GIFCT: how to handle TVEC when the role of specific groups is ambiguous, loose, or non-existent. Drawing in part on its unique dataset of TVEC, ISD's analysis of five case studies finds that, despite the heavy social media consumption and technology use of many of the attackers, many of the most important recent attacks do not directly involve specific groups. To expand the relevance of GIFCT and the HSDB in these cases, they propose a content-focused approach that differentiates "inspirational," "ideological," and "instructional" forms of TVEC, focusing on the primary function of and intent behind a given piece of content as a way to determine how to categorize it. A particular challenge is that much of the inspirational TVEC the report identifies is non-violent, such as racist memes and white-supremacist music as well the writings of ideologues like Abdallah Azzam and dystopian novels like The Turner Diaries. The authors propose a range of measures such as high-risk content flags and analyzing the broader activities of communities as possible ways to determine when to block and when to allow such non-violent but inspirational material.

As with the ISD paper, a team of authors drawn from GIFCT's Independent Advisory Committee also put together a taxonomy for content.[46] Although the scope of the paper extends beyond the HSDB itself, it offers valuable insights into how the HSDB taxonomy might be structured. By arguing for a taxonomy that matches platform type with content function, the IAC authors provide an invaluable resource for understanding the linkages between specific platform affordances and specific types of content. Much like the ISD's taxonomy, the IAC paper also offers useful distinctions between online material that contains "calls to action," consists of "ideological/strategic content," provides "material support," or facilitates "recruitment, mobilization, and retention" of group members. A key insight of the paper is that TVEC is not limited to content that simply promotes or inspires terrorist and extremist violence but also consists of content that coordinates attacks and other operational activity.

The final two papers in the series focus on specific features or aspects of a potential HSDB. Erin Miller of the University of Maryland's START program details the challenges facing the team running the Global Terrorism Database (GTD), a leading academic source of terrorism information. The GTD provides a wide range of information on terrorist events, including the actor, tactics, symbols linked to the attack, and other important data. In addition, the GTD is agnostic as to ideology, including political violence from a wide range of sources. The paper focuses on two of the knottiest problems for GIFCT when it seeks to employ academic expertise like that of START: sustainability and timeliness. Sustainability is difficult because academic institutions do not have steady funding sources despite the cost of maintaining a near real-time database, and various funding models all have numerous weaknesses. The GTD also currently has a one-year lag in its data, making it ill-suited for helping GIFCT manage real-time events. GTD experts have experimented with ways to enter events into the database with a lag of only a few days, often by triaging news articles and creating less complete and more tentative records that can then be updated. Miller argues that the GTD's experiences suggest that any expansion of the HSDB should maximize simplicity and recognize that broader inclusion of content may lead to

[46] Adam Hadley, the director of Tech Against Terrorism, was also part of this project.

more ambiguity about coding.

Finally, experts at Hedayah offer insights into how technology companies should manage the numerous definitions of terrorism and violent extremism that different governments employ. They identify common patterns among different definitions and assess their various strengths and weaknesses. Although themes like "act of violence" are near-universal, others such as "against constitutional values" are less common, while components of violent extremism vary even more. Given the wide variance, the experts recommend that technology companies develop their own definitions. However, to ensure harmony with governments, companies should employ the most common components of definitions governments use and facilitate efforts to ensure their definitions facilitate efforts to reduce extremism. Companies should also emphasize "threat, incitement, and intimidation" in their definitions due to the significant role social media can play in these factors.

Framework and Recommendations

Although the existing taxonomy for the HSDB is imperfect, improving on it is far from straightforward. As several of the papers in this series have demonstrated, the boundaries between TVEC and hate speech, and between terrorist organizations and non-violent fringe groups, belie easy demarcation. Identifying which categories of content should be hashed and shared in the HSDB will always be difficult.

One problem is simply defining terrorism. As the authors have noted in a separate paper, this is a fraught issue, with politics shaping the results as well as a range of analytic concerns.[47] However, despite considerable variation, common elements of many governmental, civil society, and academic definitions of terrorism include the following:

1. the use or threat of violence;
2. the target is or includes non-combatants;
3. the political nature of the act;
4. the violence performed by a non-state actor (thus excluding efforts by state intelligence operatives and paramilitary forces); and
5. the violence is intended to have a broader psychological effect.

Violent extremism can be seen as a broader category, where the violence need not be intended to have a psychological effect, thus including more prosaic (but by no means less deadly) attacks on enemy communities, property destruction, and so on.[48] In their comprehensive look at different government definitions of terrorism and violent extremism, the Hedayah report discussed above found wide variation in the components cited.

[47] Meserole and Byman, "Terrorist Definitions."

[48] It is important to note that groups can be "extreme" without being violent. As J. M. Berger notes, "Violent extremism is the belief that an in-group's success or survival can never be separated from the need for violent action against an out-group." J.M. Berger, Extremism (Cambridge, MA: The MIT Press, 2018), 46.

Despite the inevitable definitional ambiguity, several clear improvements to the HSDB are nonetheless possible. Drawing on insights from the commissioned paper series as well as our own research, we lay out a potential framework and set of recommendations below. Our suggestions aim to address the two biggest concerns identified above with the existing HSDB – namely, that it is more focused on actors rather than content, and that there is an ideological bias shaping the actors it includes – in a way that is both nuanced and actionable. The goal is not to construct a taxonomy in full detail, but instead to sketch a clear and feasible path forward that can be expanded on in the months and years to come.

Content Taxonomy

The primary taxonomy we propose can be found in Table 1.0. Since the HSDB consists of hashed content, our taxonomy accordingly focuses on different tiers of potential content. The tiers are ordered in terms of the severity and certainty of the potential harms and risks posed by the content they reference.

Table 1.0 Content Taxonomy

Level	Content	Examples	Challenges	Process
Tier One	Depiction or coordination of terrorist or extremist violence	Graphic video of beheadings, lynchings, bombings, etc., or attack planning	Depictions of criminal violence may look similar Depictions of T/E violence can be newsworthy Depictions of T/E violence can have legal value	Context optional Appeals process Media-protected list Protected access
Tier Two	Promotion of or calls for terrorist or extremist violence	Calls for future attacks, glorification of past violence; branded instructions for the use of violence (e.g. bomb-making material)	Calls for T/E violence may be satirical Calls for criminal violence may sound similar Calls for T/E violence can be newsworthy Calls for T/E violence can have legal value	Context optional Appeals process Media-protected list Protected access
Tier Three	Promotion of terrorist or extremist ideologies	Manifestos, speeches, terrorist publications, etc. that promote the political project of a terrorist group or extremist movement	Free speech concerns Overlap of T/E and mainstream ideologies T/E ideologies have academic or public value	Context required Appeals process Media-protected lists Protected access
Tier Four	Promotion of hateful or dehumanizing ideas	Speech or images that instill fear or loathing toward an out-group (e.g., "fear of Muslims is rational"); Memes (e.g. a frog) or symbols (e.g. a noose) that with additional context would be seen has hateful	Free speech concerns Overlap of hateful / dehumanizing and mainstream political ideas Variance in cultural norms on sensitive issues	Context required Appeals process Media-protected list Protected access

At the top of the table, in tier one, are two forms of content. The first are graphic depictions of actual terrorist or extremist violence, such as the Kasasbeh video we noted at the outset, that were used by a perpetrator, accomplice, or sympathizer. This would include not only content tied to groups like the Kasasbeh video but also graphic depictions of similar violence but of unknown origin. Since there is no uncertainty that the Kasasbeh video depicts an act of terrorist or extremist violence – i.e. even without contextual information, the depiction of an individual burned alive in a cage surrounded by onlookers is enough to classify it as TVEC – the video should clearly be included in the HSDB. By contrast, the other form of content in this tier concerns attack planning, such as content that clearly serves to coordinate an attack. The specificity of such content means there is less uncertainty about whether the content will lead to real-world harm and violence.

The second tier, meanwhile, references the promotion of terrorist or extremist violence. This can include generic calls for violence ("Kill all non-whites") as well as the glorification of prior terrorist or extremist attacks. Since this content either glorifies or seeks to inspire terrorist or other violence explicitly, the risks it poses are significant enough to merit inclusion in the HSDB, even without direct affiliation to designated terrorist or extremist groups. For example, an anonymous manifesto that defends white supremacy and urges violence against non-whites should qualify for inclusion in the HSDB, even if its provenance is unknown and it contains no references to designated individuals or actors.

The third tier, by contrast, presents content that refers to terrorist and extremist ideas and ideologies, but does not explicitly depict, promote, or call for violence. An example would be a manifesto that calls for an ethnonationalist state or a society that is racially pure but does not discuss whether violence would be required to achieve that vision. Since the content is not necessarily calling for violence, it therefore poses less of a potential harm than content in tier two.

The fourth and final tier includes content that is hateful or dehumanizing but does not specifically reference a broader terrorist or extremist ideology nor call for violence explicitly. An example would be a tweet by Michael Flynn, a former high-level official in the Trump administration, that claimed "Fear of Muslims is rational." By positioning all Muslims as threats to personal and public safety, the tweet can be read as hateful. But it does not explicitly advocate for violence against Muslims or support an extremist ideology that seeks to exclude them. Without further context, the potential harm and risk posed by such content is therefore lower. Further, the likelihood that the views overlap with legitimate mainstream political beliefs is higher. A large amount of racist, antisemitic, misogynistic, and other hateful postings and memes would fall into this category as hateful but having no clear link to immediate violence.

Several broader points about the content taxonomy are worth flagging. First, as we discuss further below, depictions and calls for terrorist or extremist violence can be newsworthy, or have legal value in documenting a crime or atrocity. GIFCT should work with member companies to architect their efforts in such a way that media-protected lists and continued access by the human rights community is supported. Second, although there are legitimate free speech concerns with every tier of the taxonomy, those concerns are more pronounced in the third and fourth tier, where the potential risk of terrorism and harm and violence is less clear and the potential overlap with mainstream partisan ideas and ideologies is greatest. As discussed below, ensuring that the HSDB has a robust and transparent appeals process will therefore be vital.

Finally, while the taxonomy provides a conceptual framework for including content in the HSDB without reference to designation lists, in many cases such lists will still be needed.

Distinguishing between depictions of or calls for criminal violence and terrorist violence, for instance, will not always be possible without cross-referencing either the creator or subject of the content with a designation list. A post calling for violence against a non-public individual, for instance, would qualify as TVEC if it came from a known violent group like the Hammerskins but not if it came from that individual's aggrieved but unaffiliated neighbor. Likewise, since content in tiers three and four contain no references to violence at all, whether it should qualify as TVEC and be included in the HSDB will always require a clear link to a designation group or individual. As a result, the HSDB will also require a taxonomy for designation lists as well.

Designation List Taxonomy

Although the HSDB should rely primarily on a content-based taxonomy, whether a piece of content qualifies for inclusion in the HSDB may also depend on contextual information about its creator or subject. More specifically, it will depend on whether the content was generated by or in support of a terrorist or extremist group or individual. Yet identifying who those groups and individuals are is itself a difficult task. Indeed, doing it well requires maintaining a list of terrorist and violent extremist actors that is objective, global in scope, and updated in real time; and even then it would be incomplete as new actors regularly emerge with little warning.

As the authors have detailed in a separate paper, using these alternative sources to construct such a list offers rich possibilities but also many problems.[49] Lists from reputable governments or organizations ostensibly offer a degree of authority, which is useful for social media companies and enables them to arbitrate without having to impose their own values – and gives them a way to deflect criticism in the process. For example, social media companies took down COVID-19-related misinformation, using the World Health Organization as an authority to categorize and justify what was true and what was false.[50] However, even if authoritarian regimes (which frequently label legitimate opposition voices as terrorists) are excluded from designation lists GIFCT members use, government lists reflect their country's political and strategic concerns, and as a result rarely overlap and are usually slow to change. Civil society organizations move quicker than governments, but their lists are often linked to a particular mission, such as combating hate or racism, and therefore include actors that are not necessarily violent. Academic lists are potentially less biased, but academics do not have the resources to act quickly or comprehensively. Lists from corporations like Facebook often draw on tremendous expertise, but corporations are not considered impartial, and the lack of transparency of much of their categorization adds to this perception. In short, there is no single source or set of sources that can easily be used to add names and organizations to the HSDB.

[49] Meserole and Byman, "Terrorist Definitions."
[50] On the value of authority, see Evelyn Douek, "Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability," Columbia Law Review, 121.1 (2020), 830.

Another drawback of actor-based lists is that they are better suited for screening content from organized groups with discrete structures rather than lone actors or leaderless movements of loosely networked individuals. Take for example white supremacists and anti-government extremists in the United States. The majority of violence involves individuals who may consume propaganda from more organized groups but largely act on their own or are in consultation with a small number of other individuals. Relying on a list of known groups, no matter how comprehensive, would not stop these individuals from posting dangerous content.

A separate and more practical concern is the challenge of consistently classifying actors across platforms. Many of the hashes submitted to GIFCT are classified by broad categories rather than actors. Different types of right-wing groups may be lumped together while individual violent actors unaffiliated with specific groups, such as the Norwegian terrorist Anders Behring Breivik, may lack their own label. Thus, for any new system to be highly effective, these companies would need to have classifying systems that are compatible with the HSDB and one another's categories.

Table 2.0 walks through a number of options for designation lists. The most limited designation list would include just the terrorist entities listed by the United Nations in its consolidated sanctions list (following existing GIFCT practice). However, as noted above, the problem with this approach is that the HSDB would remain overwhelmingly biased toward the Islamic State, Al-Qaeda, the Taliban, and their affiliate groups.

As many of the companies are U.S.-based and must comply with U.S. law, another option is to include groups on the U.S. government list of designated Foreign Terrorist Organizations and Specially Designated Global Terrorists. However, the U.S. list is an extension of U.S. foreign policy and reflects U.S. interests and politics. Technology companies would rightly be criticized for focusing too much on U.S. concerns and not enough on threats facing other countries.

By contrast, a broader government designation list would include terrorist groups designated by countries with robust civil liberties, protections for freedom of expression, and the rule of law. In theory, this would increase the perceived legitimacy of the list, and it would also allow for the inclusion of more far-right and other non-Islamist groups (since the United Kingdom, Canada, and others have recently included more such groups). However, it would also pose the difficult question of which democracies to include, and would also not be updated regularly.

A final option would be an expanded designation list that draws on democratic designation lists, but also includes groups and individuals vetted by researchers and experts from industry, academia, and civil society. No such expanded designation list is currently publicly available, although there are a number of non-governmental lists, such as those compiled by academic projects like the GTD, by companies like Facebook and Google, and by civil society actors like the Southern Poverty Law Center. Each of those lists individually contains limitations. An expanded list that draws from each should be possible to construct, though determining whom to include and the correct process requires careful consideration.

Table 2.0 Designation List Taxonomy

Content	Example	Challenges	Process
Limited Designation List	U.N. list +	Limited number of groups Focus on select Islamist groups Many causes more "individuals" and "networks"	Existing HSDB system
Broader but Select Designation List	U.N. list + U.S. list + select democratic government lists	Governments likely to lobby to include legitimate opposition groups More focus on Islamist groups than other types Many causes more "individuals" and "networks"	Establish benchmark for "democratic" Need entity to determine which governments are used List of dangerous individuals also necessary
Expanded Designation List	U.N. list + U.S. and select democratic government lists + select company and civil society lists	Governments likely to lobby to include legitimate opposition groups Many causes more "individuals" and "networks" Civil society groups' lists often vary in quality, include hateful but non-violent organizations Larger circle of governments and organizations may slow down decision making	Establish benchmark for "democratic" Need entity to determine which entities are used List of dangerous individuals also necessary

Main Recommendations for Expanding the Hash-Sharing Database

In line with the framework we describe above, our main recommendations are twofold.

First, GIFCT should use its convening power and the value of the HSDB to encourage GIFCT members to increase the standardization of their terms of service around a shared or at least compatible taxonomy of TVEC. All GIFCT members already ban content under their own terms of service or community guidelines, prohibiting content that contains, glorifies, or incites violence. This is quite broad language that would cover a large amount of content. (See Appendix A for the terms of service of GIFCT member companies.) By itself, this step would reap significant cross-platform efficiency gains without significantly changing the policies of most platforms. In addition, the platforms could learn from one another, both about differences among the platforms but also the reasoning behind how different companies are tackling the same problem. This process would be particularly useful for newer and smaller companies that have less experience and expertise in TVEC. Finally, collective action by platforms would also help insulate them from political pressure in especially controversial cases. However, increased standardization would require a significant investment on the part of the companies and considerable dialogue among them.

Second, GIFCT should work with its member companies, academia, and civil society to develop an expanded designation list or actor database. This is not a recommendation made lightly. Producing such a list would require a significant investment of time and resources, and would introduce a range of potential reputational, legal, and security risks for GIFCT. Yet the risks of not developing such a list are greater. Even with a revised content taxonomy, if the HSDB continues to also rely on the U.N.-list and/or the designation lists of established democracies, the hashes it contains will continue to reflect the political, ideological, and regional biases that inform those lists. Rather than focus overwhelmingly on Islamist extremist groups like the Islamic State or Al-Qaeda, the ideal list would instead encompass the full range of terrorist and violent extremist ideologies, including white-supremacist, far-left, and ethnonationalist ideologies. Likewise, it would also encompass extremist groups across Africa, Latin America, East Asia, and other regions that receive less global media attention than those in Europe, North America, and the greater Middle East.

Significantly, GIFCT would not necessarily need to create and maintain the list on its own but could partner with other organizations and entities to do so. Coupled with a more equitable and transparent designation process, a partnership model could increase the legitimacy of the list while also reducing its associated risks. The HSDB would become more effective as a result.

Process Implications for Expanding the Hash-Sharing Database

As we noted in the introduction, one of the main challenges related to the HSDB is not just articulating common criteria for what belongs in the database but also a common process for managing it. Although this paper is primarily focused on the criteria and taxonomy for entering content into the HSDB, the framework we have proposed nonetheless has important implications for how the HSDB should be constructed and run and how member companies should manage their own efforts. Especially since the HSDB only includes hashes rather than the underlying content, safeguards that ensure the database does not unduly restrict freedom of speech and expression globally will be essential. Some of these efforts should be led by GIFCT, but in other cases GIFCT would flag possible problems, share best practices, and otherwise assist member companies as needed should they decide to make changes. Not all of these steps can be done immediately, and indeed proper sequencing is vital.

At a minimum, those processes and safeguards should include:

Transparency reporting. For the HSDB to operate effectively, there will need to be greater transparency into how it functions and operates. The HSDB should regularly issue transparency reports that document any changes in the kinds of content the HSDB contains, as well as the processes used to collect and review that content. Any use of "tiers" or similar categories that we and several other papers in the series propose will have gray areas and coding ambiguities, and transparency will help reduce errors and create more trust in the system. Having a broad array of actors help design the categories would also assist with transparency.

Appeals process for hashed content. At present, the HSDB has an option by which companies can flag hashes that they feel were erroneously entered into the database. However, it is up to companies to check whether or not the content they have submitted has been flagged by another platform. Especially as more content is included in the database, GIFCT should work with member companies to develop a more robust notification and remediation process if two companies dispute whether a hash should be in the database.

Appeals process for designated actors. If the HSDB formalizes the designation list that member companies should use for determining whether some content belongs in the database, then it should also work with member companies to establish a process for clearly adding and removing actors from that list in a timely way. When an actor is added, the criteria should be clear regarding what it would take for the actor to be removed in terms of behavior, the length of time involved before another assessment is made, and so on.

Time-limited content. Some banned content might also be automatically time-limited, allowing for later review by companies and appeal as more information is gathered. Such an approach would allow the HSDB to be more effective in crises, as problematic content could be included, knowing a more fine-tuned approach would be applied later as companies' knowledge about the events and actors grows, and other interested parties such as academics and civil society weigh in. Time-limiting content would require a robust appeals process, however.

Regular audits. GIFCT should work with member companies and participants in the HSDB to develop and carry out regular audits of the hashes in the database. Without regular audits, there will be no way of knowing whether there are biases in the database or if it contains content that was entered mistakenly. Ideally, these audits would be undertaken by independent experts from a range of stakeholders to ensure multiple perspectives are considered and to improve credibility.

Newsworthy, academic, and legal exceptions. Although the goal of the HSDB is to limit the spread and proliferation of TVEC, some TVEC material may have public interest, research, or documentary value. GIFCT should provide guidance and work with member companies that use the HSDB to develop consistent mechanisms and processes for managing those exceptions, such as identifying media-protected lists or establishing repositories with protected access that human rights organizations or academics might access.

GIFCT should move forward to encourage more hashing of TVEC content and expanding the lists of groups beyond the current U.N.-designated violent Islamist groups. However, because the hashing of TVEC is such a powerful tool in the fight against terrorism, it must be done carefully. Hashing can be overused, interfering with a range of legitimate speech, and has the potential for mistakes. As GIFCT moves forward, it must consider the above process steps to ensure any hashing expansion is both transparent and fair.

Conclusion

The HSDB was designed to facilitate cross-program removal of TVEC. The core challenge involved is establishing common criteria for defining TVEC and a process for including it. We have proposed concrete ways of developing common taxonomy for content and actors, but implementation and process challenges remain.

Change, however, will not – and should not – happen overnight. An effective effort would involve consultation with multiple stakeholders, managing the significant technical changes for both the HSDB and with interested member companies, and include other steps that will take time. In addition, expansions to the list and additions of content may be best done in phases, beginning with the most egregious and dangerous groups and materials and, as lessons are learned, expanding the circle to less severe threats. However, given the challenge of terrorism and violent extremism, this process must begin soon.

Appendix A: GIFCT Member Companies

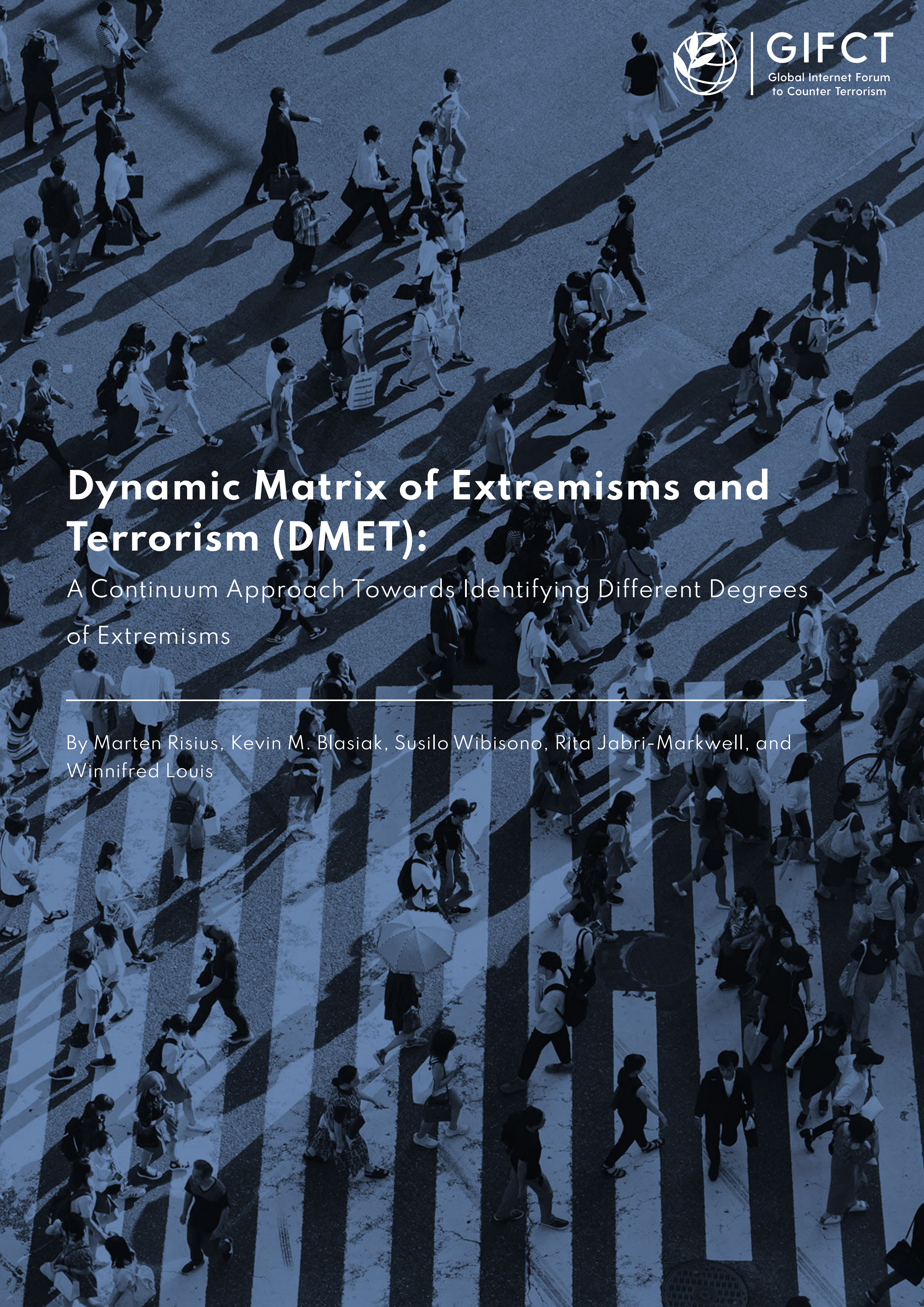
GIFCT Member Company	Company policies which prohibit content that contains, glorifies, or incites violence
Facebook	Community Standards: Violence and Incitement
Twitter	Glorification of Violence Policy
YouTube	Violent or graphic content policies
Microsoft	Microsoft Services Agreement: Code of Conduct
Mailchimp	Standard Terms of Use: Rules and Abuse
Discord	Discord Community Guidelines
Instagram	Community Guidelines
WhatsApp	WhatsApp Terms of Service: Legal and Acceptable Use
Pinterest	Community Guidelines
Amazon	Community Guidelines: Illegal Activity
Dropbox	Dropbox Acceptable Use Policy
MEGA	Mega Limited Terms of Service *
LinkedIn	Professional Community Policies
Airbnb	Community Standards **
WordPress	User Guidelines
Tumblr	Community Guidelines
JustPaste.It	N/A ***

* MEGA does not explicitly prohibit violent content, but does prohibit use of service "to abuse, defame, threaten, stalk or harass anyone, or to harm them as defined in the Harmful Digital Communications Act 2015 (NZ) or any similar law in any jurisdiction."

** Airbnb does not specifically ban content that contains or promotes violence, but it prohibits users of Airbnb to engage in violence offline.

*** JustPaste.It only contains written terms prohibiting content that is "unlawful for you to possess," including terrorism content. More found in Terms of Service, Section 7.6.

GIFCT
Global Internet Forum
to Counter Terrorism



Dynamic Matrix of Extremisms and Terrorism (DMET):

A Continuum Approach Towards Identifying Different Degrees of Extremisms

By Marten Risius, Kevin M. Blasiak, Susilo Wibisono, Rita Jabri-Markwell, and Winnifred Louis

Dynamic Matrix of Extremisms and Terrorism (DMET)

A Continuum Approach Towards Identifying Different Degrees of Extremisms

By Marten Risius, Kevin M. Blasiak, Susilo Wibisono, Rita Jabri-Markwell and Winnifred Louis

Abstract

We propose to extend the current binary understanding of terrorism (versus non-terrorism) with a Dynamic Matrix of Extremisms and Terrorism (DMET). DMET considers the whole ecosystem of content and actors that can contribute to a continuum of extremism (e.g., right-wing, left-wing, religious, separatist, single-issue). It organizes levels of extremisms by varying degrees of ideological engagement and the presence of violence identified (e.g., partisan, fringe, violent extremism, terrorism) based on cognitive and behavioral cues and group dynamics. DMET is globally applicable due to its comprehensive conceptualization of the levels of extremisms. It is also dynamic, enabling iterative mapping with the region- and time-specific classifications of extremist actors. Once global actors recognize DMET types and their distinct characteristics, they can comprehensively analyze the profiles of extremist actors (e.g., individuals, groups, movements), track these respective actors and their activities (e.g., social media content) over time, and launch targeted counter activities (e.g. de-platforming, content moderation, or redirects to targeted CVE narratives).

Key recommendations

1. Understand extremism as a dimensional concept with terrorism as a deviant pole from the regional norm.
2. Transparently map cues of extremism to accountably define categories of extremisms, creating an opportunity for dialogue between scholars and industry, and increasing trust with civil society and broader public audiences.
3. Provide the opportunity for organizations to upweigh or downweigh cues of extremism based on their local norms or national or international legal requirements, while changing the classification in an explainable and transparent way.
4. Iteratively update the manifestations of cues that characterize extremisms to account for changing profiles of extremism regionally or temporally.
5. Recognize all forms of violence used by violent extremists, especially serial or systematic dehumanization of an out-group as an attribute and indicator of violent extremism.
6. Enable platform providers to transparently decide and explain decisions to exempt extremist actors or content from DMET.
7. Create a more nuanced understanding of the degrees of extremism to reduce the probability of misclassifications (i.e., of non-terrorists as terrorists, or failing to identify terrorists as such) and allow more fine-grained analysis of actor changes over time.

Introduction

Moderation of terrorists online is commonly achieved by referring to lists of known terrorist individuals and groups from academia, civil society, and governments.[51] These lists are very helpful for the differentiation between terrorist and non-terrorist content. Currently, the GIFCT hash-sharing database is an important tool for the list-based moderation of terrorist content online.

There are, however, certain issues that accompany these list-based approaches that remain to be solved. As summarized by the recent report from the Royal United Services Institute, no single type of list simultaneously can fulfill all of the following three criteria while still being economically feasible:[52]

1. **Ideological fairness:** equal opportunity for all entities to be classified as terrorist;
2. **Global applicability:** transcend regional borders; and
3. **Update frequency:** near real-time updates.

Furthermore, there is a gray area that poses various noteworthy delicate challenges. Terrorists often communicate non-violent content that falls outside the categories of **Imminent Credible Threat, Graphic Violence Against Defenseless People, Glorification of Terror Acts, Recruitment & Instruction.** These more subtle messages still help to further extremist causes when they are not directly captured by the hash database taxonomy. They use social media for fundraising purposes[53] or to affirm grievances, ideologies, and share humanitarian purposes (e.g., pictures of ISIS-affiliated doctors helping injured children) without calling for violence.[54] Furthermore, perceived overlap in content between violent extremists and (political) partisan actors, self-determination-based movements, or state-sponsored information campaigns raise legal and ethical questions in regard to appropriate treatment. Accordingly, the decision to add an actor and their content to a list of known terrorists in order to moderate their online presence sets a high bar that allows for considerable damage to occur beforehand, is associated with a strong stigmatization transforming the decision into a political issue and makes a revision of the decision following resocialization efforts unlikely. A transparent framework like the proposed Dynamic Matrix of Extremisms and Terrorism (DMET) is needed to respond to these challenges.

[51] Chris Meserole and Daniel Byman, "Terrorist Definitions and Designations Lists: What Technology Companies Need to Know," Royal United Services Institute for Defence and Security Studies, (July 2019), https://rusi.org/explore-our-research/publications/special-resources/terrorist-definitions-and-designations-lists-what-technology-companies-need-to-know.

[52] Meserole and Byman, "Terrorist Definitions and Designations Lists," 2.

[53] Tom Keatinge and Florence Keen, "Social Media and (Counter) Terrorist Finance: A Fund-Raising and Disruption Tool," Studies in Conflict & Terrorism 42, no. 1-2 (2019); Tom Keatinge, Florence Keen, and Kayla Izenman, "Fundraising for Right-Wing Extremist Movements," The RUSI Journal 164, no. 2 (2019).

[54] Roderick Graham, "Inter-Ideological Mingling: White Extremist Ideology Entering the Mainstream on Twitter," Sociological Spectrum 36, no. 1 (2016).

Background

The Dynamic Matrix of Extremisms and Terrorism (DMET)

We propose extending the binary understanding of terrorism (versus non-terrorism) with a Dynamic Matrix of Extremisms and Terrorism (DMET) to address the intersection of extremism types (and associated extremist content). DMET identifies varying degrees of ideological engagement and violence based on cognitive and behavioral cues as well as group dynamics. [55]

DMET's Understanding of Extremism

DMET understands extremism on a continuum of varying degrees of ideological engagement. Any label of "extremism" assigned in reference to DMET needs to share the spirit of the following assumptions that accompany DMET's continuum-based understanding.

First, the associated types of extremisms are based on an understanding of online extremism as a deviation from something that is commonly considered (more) "ordinary," "mainstream," or "normal."[56] DMET declines an evaluative notion of the purposes and goals of the different forms of extremism.

Second, the levels of ideological engagement are derived from an understanding of radicalization as the "change in beliefs, feelings, and behaviors in directions that increasingly justify intergroup violence and demand sacrifice in defense of the group."[57] Consequently, we focus on cognitive, behavioral, and group dynamic cues to describe the different levels of ideological engagement. These general descriptions then need to be operationalized respective to the regional and temporal context.

Third, DMET emphasizes the plurality of extremisms to underscore our assumption that extremism is a concept of varying degrees and deviation from regionally dominant ideologies. We emphasize this to avoid stigmatization of minorities as "extremists" for proposing views that deviate from the regional majority (e.g., Radical Veganism). Higher

[55] Peter R. Neumann, "The Trouble with Radicalization," International Affairs 89, no. 4 (2013); Kris Christmann, "Preventing Religious Radicalisation and Violent Extremism: A Systematic Review of the Research Evidence," Youth Justice Board (2012); Susilo Wibisono, Winnifred R Louis, and Jolanda Jetten, "A Multidimensional Analysis of Religious Extremism," Frontiers in Psychology, 10 (2019).

[56] Alex P. Schmid, "Violent and Non-Violent Extremism: Two Sides of the Same Coin," International Centre for Counter-Terrorism (ICCT) Research Paper (2014); Charlie Winter et al., "Online Extremism: Research Trends in Internet Activism, Radicalization, and Counter-Strategies," International Journal of Conflict and Violence (IJCV) 14, no. 2 (2020): 4; Ronald Wintrobe, Rational Extremism: The Political Economy of Radicalism, Cambridge, New York, Cambridge University Press, (2006).

[57] Clark McCauley and Sophia Moskalenko, "Mechanisms of Political Radicalization: Pathways toward Terrorism," Terrorism and Political Violence 20, no. 3 (2008); Winter et al., "Online Extremism."

levels of ideological engagement in different forms of extremisms are characterized by more uncommon cognitive, behavioral, and group dynamic cues.

Fourth, we consider the matrix to be dynamic to acknowledge both that groups change and that the understanding of "normal" is variable across geography and time. For example, a group that opposes vaccination may be viewed as fringe or extremist (i.e., non-normative) in certain parts of the world at a particular time, but regarded as normal in other parts or at other times.[58] Hence, assessments of ideological engagement and the underlying forms of operationalizations are regionally delimited, time-specific, and require regular updates.

Level Defining Cues

DMET distinguishes between four levels of ideological engagement: partisan, fringe, violent extremist, and terrorist. DMET's continuum approach adopts the idea of ordering degrees of violent extremism[59] to extend the simplified categories of terrorism and non-terrorism. In DMET's case, we start from a point at the normative or moderate baseline and move through to increasing degrees of alienation from the mainstream to ultimately active violent acts. Each level of ideological engagement is proposed to have a particularly prevalent configuration of cues to identify and classify a group or content (i.e., from partisanship to terrorism). These cues are cognitive, behavioral, and group dynamic. A discussion of strategic and technical implementation considerations can be found in sections 4.2 and 4.3.

Cognitive Cues in DMET

Cognitive cues are signals indicating the thoughts and attitudes of individuals or groups.[60] At the individual level, cognitive cues might emerge in the form of thoughts or images. At the collective or group level, cognitive cues are shared beliefs or representations involved in recognizing and perceiving ourselves and other individuals or social categories. Many outcomes can flow from these socio-cognitive processes, such as prejudice and stereotypes.[61] For the purpose of classifying content as extremist or not, key aspects tracked by DMET cognitively would include beliefs about or representations of one's own groups (ingroups) and their actions, the targeted opponent groups (out-groups) and their actions, the nature of right or wrong, and the nature of the threats or value differences that define the relationship between the groups.

[58] Ayodele Samuel Jegede, "What Led to the Nigerian Boycott of the Polio Vaccination Campaign?," PLoS Medicine 4, no. 3 (2007).

[59] For an example, see Donald Holbrook, "Designing and Applying an 'Extremist Media Index,'" Perspectives on Terrorism 9, no. 5 (2015).

[60] Bert N. Bakker, Yphtach Lelkes, and Ariel Malka, "Understanding Partisan Cue Receptivity: Tests of Predictions from the Bounded Rationality and Expressive Utility Perspectives," The Journal of Politics 82, no. 3 (2020).

[61] M. Verkuyten and A. De Wolf, "The Development of in-Group Favoritism: Between Social Reality and Group Identity," Developmental Psychology 43, no. 4 (2007).

Behavioral Cues in DMET

Behavioral cues refer to the observable actions by groups or individuals or representations of those actions. At an individual level, behavioral cues can be observed from (for example) facial expression, gesture, vocal expression, etc.[62] For the purposes of DMET, these cues might indirectly identify (for example) the emotional level (e.g., anger) that a person has in one situation, or may explicitly show harm-doing and calls to violence. Behavioral cues can be addressed to the self or others. At a group level, drawing on the literature on political contestation, collective action,[63] and intergroup violence. For the purposes of DMET, we are most interested in coding for content that involves a call to cooperate with in-group or prospective allies or to engage in concrete actions that derogate or harm another group.

Group Dynamics in DMET

The process of radicalization or increasing extremism often draws on group dynamics by establishing norms about appropriate and deviant behaviors, with very little latitude in accepting differences.[64] People may be drawn to identify with causes or groups based on broad in-/out-group dynamics, as intergroup threats and conflicts of interest or values are contested using a range of tactics, from debate and satire to threats, dehumanization, and violence.[65] The dynamic influence of group identities and norms can provide ideological glue for (de)radicalization across the extremist spectrum.[66]

Group dynamics refers to a system of behaviors and psychological processes occurring within or between social groups.[67] Intragroup dynamics (i.e., how individuals in a group interact with one another) underlie social processes that give rise to a set of norms, roles, relations, and common goals characterizing a particular group.[68] Group dynamics can also involve the cooperation or competition of individuals within the groups to gain group recognition or act on behalf of the group. In addition, intergroup dynamics (i.e., how groups interact with each other) include collective perception, attitudes, and actions

[62] Alessandro Vinciarelli et al., "Social Signal Processing: State-of-the-Art and Future Perspectives of an Emerging Domain," in Proceedings of the 16th ACM international conference on Multimedia (Vancouver, British Columbia, Canada: Association for Computing Machinery, 2008).
[63] Defined as any action aimed to improve the group's status; see M. van Zomeren, T. Postmes, and R. Spears, "Toward an Integrative Social Identity Model of Collective Action: A Quantitative Research Synthesis of Three Socio-Psychological Perspectives," Psychological Bulletin 134, no. 4 (2008); S. C. Wright, D. M. Taylor, and F. M. Moghaddam, "Responding to Membership in a Disadvantaged Group - from Acceptance to Collective Protest," Journal of Personality and Social Psychology 58, no. 6 (1990).
[64] Wibisono, Louis, and Jetten, "A Multidimensional Analysis of Religious Extremism."
[65] C. Stott, P. Hutchison, and J. Drury, "'Hooligans' Abroad? Inter-Group Dynamics, Social Identity and Participation in Collective 'Disorder' at the 1998 World Cup Finals," British Journal of Social Psychology, 40 (2001).
[66] John M. Berger, "Deconstruction of Identity Concepts in Islamic State Propaganda: A Linkage-Based Approach to Counter-Terrorism Strategic Communications," The Hague, Netherlands: EUROPOL, (2017); Donald Holbrook, "Far Right and Islamist Extremist Discourses: Shifting Patterns of Enmity," Extreme Right Wing Political Violence and Terrorism (2013).
[67] M. A. Hogg and D. J. Terry, "Social Identity and Self-Categorization Processes in Organizational Contexts," Academy of Management Review 25, no. 1 (2000); J. Sidanius et al., "Ethnic Enclaves and the Dynamics of Social Identity on the College Campus: The Good, the Bad, and the Ugly," Journal of Personality and Social Psychology 87, no. 1 (2004).
[68] M. A. Hogg and S. A. Reid, "Social Identity, Self-Categorization, and the Communication of Group Norms," Communication Theory 16, no. 1 (2006).

toward other groups.[69]

Intra- and intergroup dynamics can produce and be shaped by specific behavioral and cognitive cues. However, the DMET coding here refers to specific attributes of how content is being disseminated relationally (e.g., conformity, polarization), and how sources are positioning themselves in relation to other groups (e.g., as leaders, warriors) and groups in relation to each other (e.g., as enemies, allies, dupes). Source attributes where available would be coded in Group Dynamics, both in terms of membership in particular groups and of position within particular networks (e.g., contact with a known violent actor).

Types of Ideological Engagement

Crossed with these levels in DMET (see Figure 1), we consider five categories of actors/content according to their ideological arena: Right-Wing (e.g., concerning threats to the "white race" or "traditional values"), Left-Wing (e.g., concerning the need for a fair distribution of wealth), Religious (e.g., seeking to spread one's religion or purify it), Separatist (e.g., seeking territory for one's group), and Single-issue (e.g., advocating for one particular topic such as abortion or animal justice).[70] A group may be classified into more than one type of ideology, as it advocates for an issue by drawing narratively on other content (e.g., both right-wing ideology and religion). The purpose of the categories is to a) signal the inclusivity of DMET with all groups equally able to be considered as violent actors or terrorists; and b) build an understanding of how clusters of particular indicators or attributes emerge in different causes, resulting in profiles of domain-specific indicators feeding into context-specific categorization algorithms.

[69] "Hooligans' Abroad?."

[70] Allard R. Feddes et al., Psychological Perspectives of Radicalization (London: Routledge, 2020).

The Continuum of Ideological Engagement

A core premise of DMET is its understanding of ideological engagement as a spectrum of varying degrees of severity (Figure 2) instead of the current binary dichotomy of (non) terrorism (Figure 1).

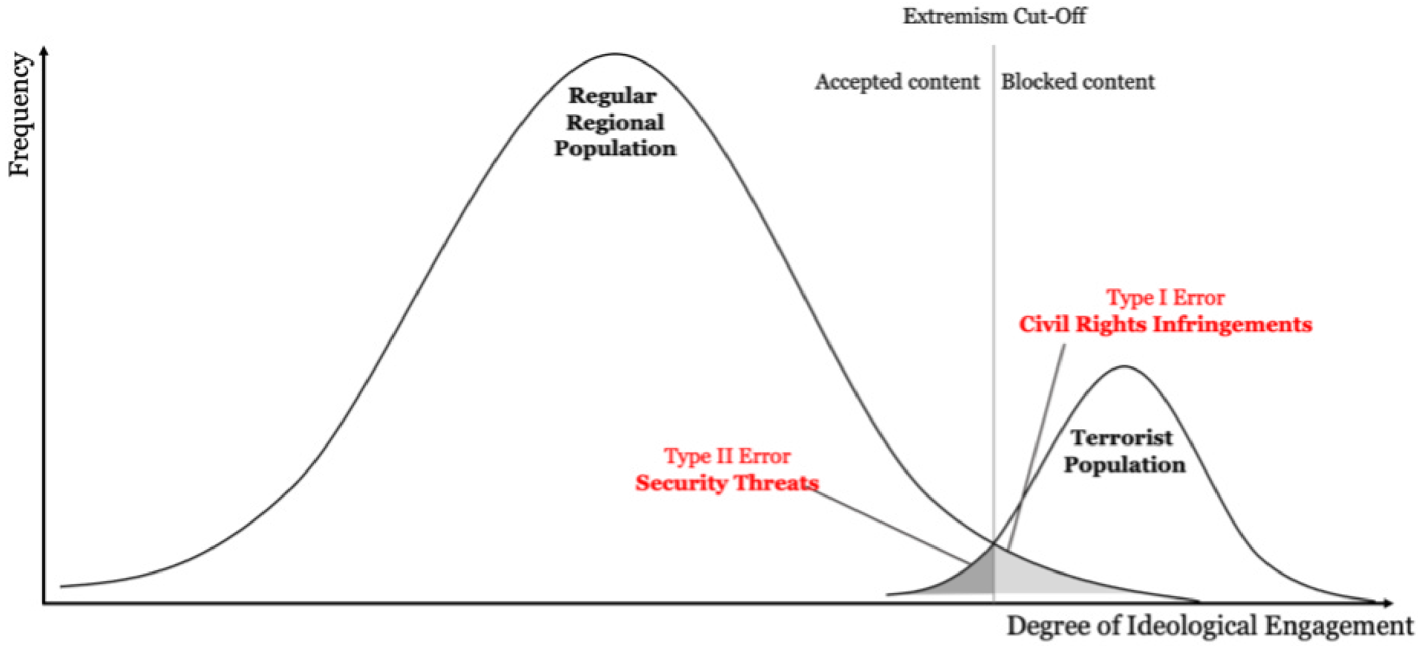


Figure 1. Conceptualization of Current Dichotomous Understanding of Ideological Engagement

This continuum perspective enables DMET to distinguish among different levels of ideological engagement (i.e., partisanship, fringe, violent extremism, terrorism) and the regular population norms that define the regionally accepted social standard. Thereby, the aim is to enable platform providers to make independent content or actor moderation decisions in a more nuanced fashion (including determining and disclosing cut-off values), with fewer misclassification errors between regular content and terrorist material. DMET also enables greater transparency regarding moderation decisions (e.g., by providing a means not only to classify organizations among the dimensions but also to develop and explain the weighing of attributes in the decision-making algorithms).

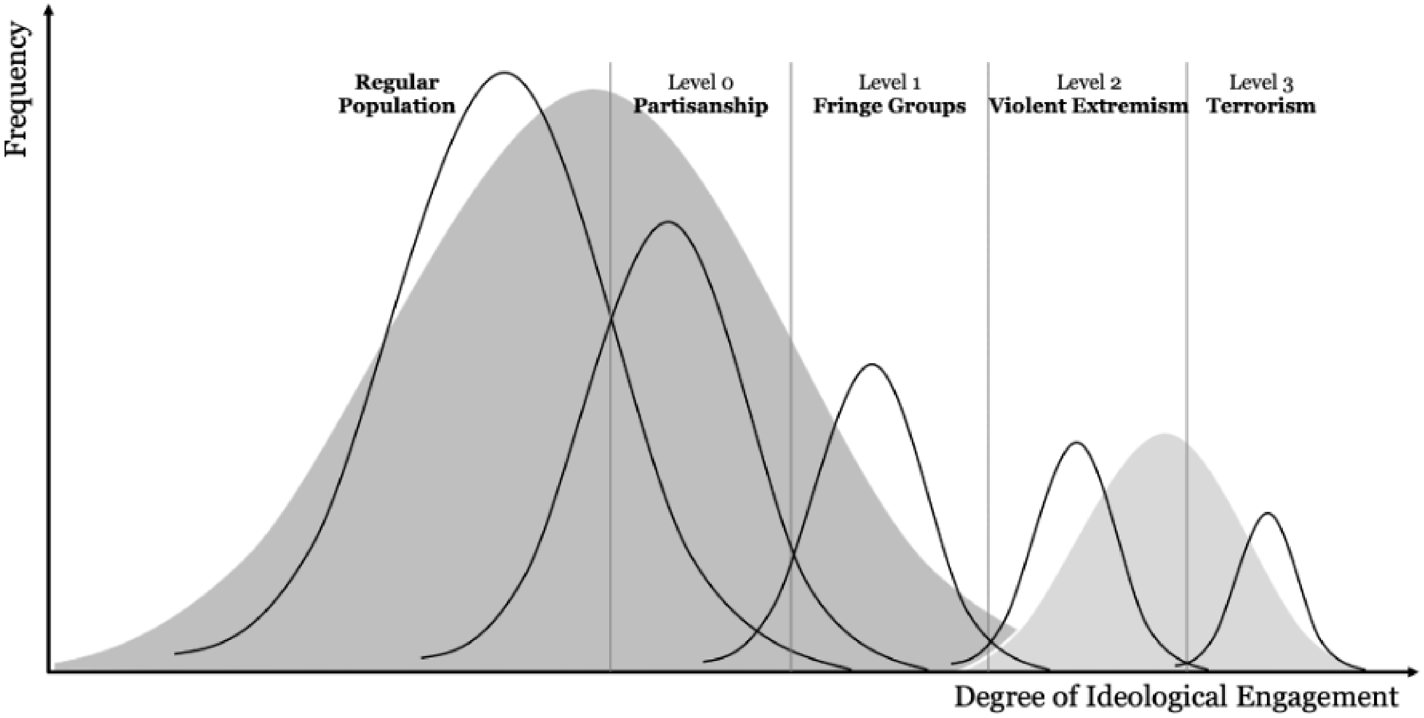


Figure 2. Conceptualization of DMET's Proposed Level's of Extremism Based on the Assumed Continuum of Ideological Engagement

It needs to be noted that the conceptualization of the different levels of ideological engagement as normally distributed sub-groups is a proposition that needs to be empirically tested following DMET's operationalization criteria outlined in the following. We also emphasize that this is meant as a conceptual illustration of the continuum of extremism and that the proportions of the more extreme sub-populations are likely smaller in reality.[71]

[71] Dirk Oegema and Bert Klandermans, "Why Social Movement Sympathizers Don't Participate: Erosion and Nonconversion of Support," American Sociological Review 59, no. 5 (1994).

Levels of Ideological Engagement	Level Defining Cues			Types of Ideological Engagement				
	Cognitive	Behavioral	Group Dynamic	Right-Wing	Left-Wing	Religious	Separatist	Single-Issue
	Terrorism							
Level 3 Terrorism	Sidestep inhibitory mechanisms, perceive target as 'the enemy," legitimizing and valorizing death, wanting to intimidate broader population	Endorse, promote, or enact physical violence towards out-, in-group or infrastructure	Propagate values of active martyrdom, divide group labor to support violent acts	Ultra-right, far-right, alt-right, right-wing extremism, fascism, white supremacy	Ultra-left, far-left, left-wing extremism	Religiously motivated terrorism	Violent militant separatist organizations	Violent militant activism
	Violent Extremism							
Level 2 Violent Extremism	Be intolerant towards others, represent cultural & structural violence through silencing and exclusion, perceive a reduced level of moral duties owed to the out-group	Serially or systematically dehumanize others, frequently express hate speech towards opponents, perform selective/ individual acts of violence, actively separate targets from society, active discrimination	Compete for within-group recognition, show personal agency in the service of group domination, coalescing around out-group as a perceived or designated existential threat	Radical right, extreme conservatism	Radical left	Religiously motivated extremism	Secessionism, autonomism	Propaganda groups
	Non-Violent Extremism							
Level 1 Fringe Group	Perceive/glorify the in-group as superior, indoctrinate dogmatic values, prejudice, and discrimination	Discredit or denigrate the out-group, seek isolation from the general public, express external blame for negative events, censor deviant views	Pursue and promote norms of purity, supremacy, domination, or revenge	Right-wing nationalist	Left-wing nationalist	Religious fundamentalism, cult	Seeking self-determination	Conspiracy theorists, fringe party advocating single-issue
	Non-Extremism							
Level 0 Partisanship	Holding polarized and normative views, self-identifying with one group in opposition to another group	Expressing populist ideology, dog-whistling, satirizing other views, evangelizing others, campaigning peacefully	Holding political grievances, experiencing a sense of victimization or identity crises, or a need for significance	Right-wing populism	Left-wing populism, liberalism, socialism	Religious conservatism	Regional advocacy groups	Special-interest advocacy groups, lobbyism

Please note: The table is based on a value-free understanding of extremism as something that is significantly deviant from the 'mainstream' or 'normal'; dynamic classifications of content, individuals or groups are dependent on the understanding of what is 'normal' in a particular region at a given point in time.

Table 1. Dynamix Matrix of Extremisms and Terrorism (DMET)

Operationalization of DMET

In the following, we describe DMET's operationalization of the proposed levels of ideological engagement based on cognitive, behavioral, and group dynamic cues. We acknowledge the wealth of literature discussing comprehensive definitions of these concepts. For the sake of this briefing paper, we limit ourselves to deriving working definitions that explain DMET-based actor classifications.

We identify an indicative basket of indicator attributes for each level. However, part of our approach is that the association of any one attribute with a level or ideological cause (e.g., right-wing extremism) is dynamic and may change over time, so indicators wax or wane in their diagnostic value in historical periods or for particular contexts. Regional expert feedback to set the starting parameters and automated updating of the model over time will be important in sustaining DMET's accuracy.

Level 0: Partisanship

Partisanship constitutes a non-extremist form of coordinated ideological engagement where individuals are committed to similar normative ideas and face conditions of conflict opposing others with whom they are at odds.[72] Partisans distinguish themselves from mainstream views through their normative ideology and offer a support network where collective actors are empowered to contest perceived grievances. The partisan commitment also serves as a source of identification and shapes the individuals' self-concept,[73] which leads to the continued endorsement of the mission that the group embodies and sustains the long-term pursuit of projects across a range of conditions and circumstances. Partisanship is not limited to in-group conformity but also is associated with perceived polarization away from a rival or opponent out-group.[74]

At a behavioral level, partisans commit to enacting a form of regulated adversarialism, which describes their commitment to persuade and evangelize others of their views tempered by self-set rules or ideals.[75] They may pursue different strategies, such as expressing populist ideologies as an opposition force or from a position of power.[76] Traditionally partisan actions take place through conventions, meetings, assemblies, and peaceful protests complemented by the online sphere via websites, blogs, and social

[72] Jonathan White and Lea Ypi, The Meaning of Partisanship, Oxford, Oxford University Press, (2016).
[73] Emily A. West and Shanto Iyengar, "Partisanship as a Social Identity: Implications for Polarization," Political Behavior (2020).
[74] Noam Lupu, "Party Polarization and Mass Partisanship: A Comparative Perspective," Political Behavior 37, no. 2 (2015).
[75] White and Ypi, The Meaning of Partisanship.
[76] S. Erdem Aytaç, Ali Çarkoğlu, and Ezgi Elçi, "Partisanship, Elite Messages, and Support for Populism in Power," European Political Science Review 13, no. 1 (2021).

media. Calls for action against the out-group are often indirect, using appeals that subtly invoke negative stereotypes about an opposing group (e.g., dog-whistling or racial priming theory) to harness the power of prejudice.[77] Partisans often target mainstream audiences, satirizing others by embedding ideological information into entertaining formats to engage others who are otherwise agnostic about a particular issue.[78]

In terms of group dynamics, partisans express and market feelings of injustice, grievances, or disaffection,[79] invoking personal and collective needs for significance and a desire to matter and be respected.[80] Narratives often identify an identity crisis threatening the group[81] and victimization at the hands of other out-groups.

Level 1: Fringe Groups

According to DMET, fringe groups describe non-violent ideologies that are on the periphery of social movements or larger organizations, with more extreme views than those of the majority. Again, we acknowledge the considerable heterogeneity of ways to be a fringe actor and also the reality that in particular contexts, the toxic dynamics we ascribe below to "fringe" organizations may also apply to mainstream partisan groups. Based on this logic, we propose that DMET platforms create the opportunity to transparently and accountably "dial down" the diagnostic weighing of a particular dimension (e.g., out-group derogation) to avoid false positives when such rhetoric characterizes mainstream discourse.

With that caveat noted, DMET proposes that fringe groups are marked by cognitive cues such as beliefs of in-group superiority, out-group distinctiveness and inferiority, dogmatic values, learned prejudice, and discrimination.[82]

Behaviorally, we conceive that fringe groups discredit or denigrate the out-group, promote isolation from the general public, and promote narratives of external blame for negative outcomes such as conspiracy theories.[83]

[77] Rachel Wetts and Robb Willer, "Who Is Called by the Dog Whistle? Experimental Evidence That Racial Resentment and Political Ideology Condition Responses to Racially Encoded Messages," Socius 5 (2019).

[78] Silvia Knobloch-Westerwick and Simon M. Lavis, "Selecting Serious or Satirical, Supporting or Stirring News? Selective Exposure to Partisan Versus Mockery News Online Videos," Journal of Communication 67, no. 1 (2017).

[79] Donald R. Kinder and D. Roderick Kiewiet, "Economic Discontent and Political Behavior: The Role of Personal Grievances and Collective Economic Judgments in Congressional Voting," American Journal of Political Science (1979); White and Ypi, The Meaning of Partisanship.

[80] Arie W. Kruglanski et al., "The Psychology of Radicalization and Deradicalization: How Significance Quest Impacts Violent Extremism," Political Psychology 35 (2014).

[81] John Sides, Michael Tesler, and Lynn Vavreck, Identity Crisis: The 2016 Presidential Campaign and the Battle for the Meaning of America (Princeton University Press, 2019).

[82] Roy F. Baumeister, Evil: Inside Human Cruelty and Violence (WH Freeman/Times Books/Henry Holt & Co, 1996); Robert J. Sternberg, "A Duplex Theory of Hate: Development and Application to Terrorism, Massacres, and Genocide," Review of General Psychology 7, no. 3 (2003).

[83] Marc W. Heerdink et al., "Emotions as Guardians of Group Norms: Expressions of Anger and Disgust Drive Inferences About Autonomy and Purity Violations," Cognition and Emotion 33, no. 3 (2019).

In turn, the group dynamics of fringe actors are marked by internal intolerance and censorship of deviant views, as well as readiness to pursue and promote norms of purity, supremacy, domination, or revenge.[84] Group members are socialized and indoctrinated into binary right–wrong classifications, sometimes in a highly systematic fashion in which newcomers move from the ideological periphery of their group to the inside through contracts of commitment and conversion, and in concert to withdrawing in isolation from other sources of identity such as family.[85]

Level 2: Violent Extremism

Violent Extremists propagate a radical ideology supported by violent means that condone physical or mental harm to others. A key definitory factor that determines violent extremism is ideologically sanctioned violence such as dehumanization. Our concept is that groups often differ internally in the tactics advocated and contest the use of violence, and we seek to distinguish fringe groups in which isolated and/or peripheral members advocate for hate or violence from violent extremist groups where leaders and mainstream advocates do so, to terrorist groups where a formal division of labor to carry out attacks has been implemented.

On a cognitive level, violent extremists are intolerant towards others, representing cultural and structural violence through silencing and exclusion as just, inevitable, or appropriate, perceiving a reduced level of moral duties owed to the out-group. Extremists develop narratives legitimizing violence, often by framing the out-group as an enemy who is violent towards them.[86]

Behaviorally, violent extremists serially or systematically dehumanize others. Violent extremists refuse to tolerate or respect opinions or beliefs contrary to their own; they perceive a moral superiority and obligation to enforce their ideology.[87] This also frees extremists to act violently against the "other" without moral obligations and the burden of guilt that would typically be associated with violence.[88] Against that backdrop, these groups

[84] Dominic Abrams et al., "Pro-Norm and Anti-Norm Deviance within and between Groups," Journal of Personality and Social Psychology 78, no. 5 (2000); Dominic Abrams et al., "Collective Deviance: Scaling up Subjective Group Dynamics to Superordinate Categories Reveals a Deviant Ingroup Protection Effect," Journal of Personality and Social Psychology (2021); Roger Giner-Sorolla, Bernhard Leidner, and Emanuele Castano, "Dehumanization, Demonization, and Morality Shifting: Paths to Moral Certainty in Extremist Violence," Extremism and the Psychology of Uncertainty (2012).

[85] Andrew Coulson, "Education and Indoctrination in the Muslim World," Policy Analysis 29 (2004); Michael A. Hogg, Arie Kruglanski, and Kees Van den Bos, "Uncertainty and the Roots of Extremism," Journal of Social Issues 69, no. 3 (2013); F. M. Moghaddam, "The Staircase to Terrorism a Psychological Exploration;" American Psychologist 60, no. 2 (2005); John G. Horgan et al., "From Cubs to Lions: A Six Stage Model of Child Socialization into the Islamic State," Studies in Conflict & Terrorism 40, no. 7 (2017).

[86] Douglas Pratt, "Religion and Terrorism: Christian Fundamentalism and Extremism," Terrorism and Political Violence 22, no. 3 (2010); David Webber and Arie W. Kruglanski, "The Social Psychological Makings of a Terrorist," Current Opinion in Psychology 19 (2018).

[87] Hogg, Kruglanski, and Van den Bos, "Uncertainty and the Roots of Extremism."

[88] Erving Goffman, Stigma: Notes on the Management of Spoiled Identity (Simon and Schuster, 2009); Albert Bandura, "Moral Disengagement in the Perpetration of Inhumanities," Personality and Social Psychology Review 3, no. 3 (1999).

frequently express hate speech towards opponents to create psychological and structural violence through silencing and exclusion. Individual members of violent extremist groups may perform selective acts of physical violence as part of a group dynamic that valorizes these actions. Hate speech and glorification of violent acts would both be indicators of this level of engagement in DMET.[89]

At the group level, members of violent extremist groups compete for within-group recognition, seeking to show personal agency in the service of group domination, and coalescing around out-groups as perceived or designated existential threats.[90] The normative context of dehumanization establishes social preconditions within which violence by extremist instigators is likely to be perceived as justified. They authorize individuals to perform violence and shape bystanders' reactions to these events, while establishing the parameters for depersonalization and stigma or dehumanization and moral exclusion.[91] While these group dynamics might not be transparent at the content level, favorable responses valorizing particular in-group actors who are violent may provide a key set of indicators that would serve to identify the dynamics at play.[92]

Level 3: Terrorism

Terrorism constitutes the most extreme form of ideologically driven engagement that uses violence even towards non-combatant targets to instill terror or to send a 'message'.[93] At the cognitive level, terrorists experience two key psychological processes involving a rigid, exclusive social categorization (e.g., of civilians as part of the out-group) and a greater psychological or moral distance by exaggerating differences between the in-group and the out-group.[94] The categorization of society at large as part of the out-group and as the enemy then serves as the justification for their struggle to intimidate or harm civilians.[95] Terrorists thereby sidestep "inhibitory mechanisms" that would normally limit the aggression of humans against one another. Instead, they show the greatest adherence to principles that move them to conform unconditionally to certain moral duties, which

[89] Alexandra Olteanu et al., "The Effect of Extremist Violence on Hateful Speech Online" (paper presented at the Proceedings of the International AAAI Conference on Web and Social Media, 2018); Pratt, "Religion and Terrorism."

[90] Gary A. Ackerman, Jun Zhuang, and Sitara Weerasuriya, "Cross-Milieu Terrorist Collaboration: Using Game Theory to Assess the Risk of a Novel Threat," Risk Analysis 37, no. 2 (2017); Randy Borum, "Radicalization into Violent Extremism I: A Review of Social Science Theories," Journal of Strategic Security 4, no. 4 (2011); David R. Mandel, "The Role of Instigators in Radicalization to Violent Extremism," Psychosocial, Organizational, and Cultural Aspects of Terrorism: Final Report to NATO HFM140/RTO. Brussels: NATO (2011).

[91] Erving Goffman, Stigma: Notes on the Management of Spoiled Identity (Simon and Schuster, 2009); Albert Bandura, "Moral Disengagement in the Perpetration of Inhumanities," Personality and Social Psychology Review 3, no. 3 (1999).

[92] Arie W. Kruglanski et al., "To the Fringe and Back: Violent Extremism and the Psychology of Deviance," American Psychologist 72, no. 3 (2017); Jeff Victoroff, Janice R. Adelman, and Miriam Matthews, "Psychological Factors Associated with Support for Suicide Bombing in the Muslim Diaspora," Political Psychology 33, no. 6 (2012); Webber and Kruglanski, "The Social Psychological Makings of a Terrorist."

[93] Meserole and Byman, "Terrorist Definitions and Designations Lists."

[94] Moghaddam, "The Staircase to Terrorism."

[95] Marc Sageman, Understanding Terror Networks (University of Pennsylvania Press, 2011).

ultimately legitimize and valorize death.[96] Cognitive beliefs about the legitimacy of killing and the glory of risking sacrificial death are often indicators of the terrorist level in DMET.

Behaviorally, terrorists also endorse, promote, and engage in violent and destructive actions. These are predominantly directed at civilians as well as non-human symbolic or infrastructure targets (e.g. works of art, places of worship).[97] Terrorists target different objectives depending on the specific sources of support available to them and the degree of out-group antagonism in their constituency.[98] Terrorists, however, also engage in violent actions against in-group members as (potential) defectors to sustain the long-term mission and group norms.[99] Concrete incitement to violence and physically violent acts provide behavioral indicators of the terrorist level in DMET.

In terms of group dynamics, terrorists have organized social structures that support violent actions on an ongoing basis. Terrorists often are taught to internalize the glorification of active martyrdom as a testimony of ideological commitment and faith.[100] We refer to active martyrdom (as opposed to passive martyrdom) as a characteristic of terrorism in the sense of a suicide attack where the act of self-destruction targets a perceived out-group enemy. Passive martyrdom (e.g., in politics or religion), where the actor is compelled to be ready to give one's life to defend the ideals and values of the group, is not considered a terrorist attribute.[101] However, while these beliefs are in theory shared by all group members, in practice, a formal division of labor to support violent acts exists (e.g., consisting of finances, military affairs, religious affairs, and public relations).[102]

In some cases, demographic divisions stream actors to different roles (e.g., younger men might be expected to serve as martyrs while older men direct actions and women serve support roles). In other cases, core groups of strategists and recruiters with ongoing roles might engage opportunistically, at a distance, with individuals of any age and gender recruited as one-off cannon fodder. These group dynamics might be discerned through representations of terrorists in the inward-facing communications of the group (e.g., distinctive costumes and language) or might be coded as attributes associated with particular sources.

[96] Domenico Tosini, "Calculated, Passionate, Pious Extremism: Beyond a Rational Choice Theory of Suicide Terrorism," Asian Journal of Social Science 38, no. 3 (2010).
[97] Pratt, "Religion and Terrorism."
[98] Sara M.T. Polo, "The Quality of Terrorist Violence: Explaining the Logic of Terrorist Target Choice," Journal of Peace Research 57, no. 2 (2020).
[99] Daniel Koehler, "Radical Groups' Social Pressure Towards Defectors: The Case of Right-Wing Extremist Groups," Perspectives on Terrorism 9, no. 6 (2015).
[100] Sageman, Understanding Terror Networks.
[101] David Cook, "The Implications of "Martyrdom Operations" for Contemporary Islam," The Journal of Religious Ethics 32, no. 1 (2004).
[102] Sageman, Understanding Terror Networks.

Primary Role of Dehumanization for Distinguishing Levels 1 and 2 (Fringe Actors Versus Violent Extremists)

The violent extremist category in DMET (Level 2) includes actors (and content) that is either associated with physical violence or associated with non-physical violence in the form of dehumanization. Facebook (and by relation Instagram), Twitter, YouTube, and LinkedIn recognize dehumanization as a particularly dangerous form of hatred as it removes moral objections one may have to enact violence, even mass violence, against women,[103] children,[104] and civilians more broadly within a target group. It connects to violent extremists' cognition of representing cultural and structural violence through silencing and exclusion. It supports their group dynamics of coalescing around an out-group as the perceived or designated existential threat. While dehumanization may not always lead to violence, genocides and atrocities typically require it. This cue would identify groups or individuals that rely on dehumanizing language, or over time are spreading large amounts of dehumanizing discourse about a group identified on the basis of a protected characteristic. Dehumanization occurs in two forms:

1. Dehumanizing language includes material that presents the class of persons to have the appearance, qualities or behavior of an animal, insect, filth, form of disease or bacteria; or to be inanimate or mechanical objects; or a supernatural threat, in circumstances in which a reasonable person would conclude that the material was intended to cause others to see that class of persons as less deserving of being protected from harm or violence. This material would include words, images, and/or insignia;[105] and
2. Dehumanizing discourse or conceptions include the sustained curation of information to a specific audience to suggest that the class of persons on the basis of their identified characteristic[106]

    a. are polluting, despoiling, or debilitating society;

    b. have a diminished capacity for human warmth and feeling or independent thought;

    c. act in concert to cause mortal harm; or

    d. are to be held responsible for and deserving of collective punishment for the specific crimes, or alleged crimes of some of their "members."

[103] Nikki Marczak, "A Century Apart: The Genocidal Enslavement of Armenian and Yazidi Women," in A Gendered Lens for Genocide Prevention, ed. Mary Michele Connellan and Christiane Fröhlich (London: Palgrave Macmillan UK, 2018).
[104] Peter Lentini, "The Australian Far-Right: An International Comparison of Fringe and Conventional Politics," in The Far-Right in Contemporary Australia, ed. Mario Peucker and Debra Smith (Singapore: Springer Singapore, 2019).
[105] Nick Haslam, "Dehumanization: An Integrative Review," Personality and Social Psychology Review 10, no. 3 (2006); Jonathan Leader Maynard and Susan Benesch, "Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention" Genocide Studies and Prevention: An International Journal 9, no. 3 (2016).
[106] Haslam, "Dehumanization: An Integrative Review"; Maynard and Benesch, "Dangerous Speech and Dangerous Ideology."

While preventing dehumanization is an imperative under international law (e.g., Article 20, 2, ICCPR; Article 25, 3e of the Rome Statute) current algorithms are focused on detecting individual instances. We conceive that DMET could be trained to predict aggregate harm by specific actors from a range of samples of borderline content that each might be difficult to discern as harmful individually. Information campaigns acting as vehicles for widespread dissemination of dehumanizing conceptions and discourse will need to be distinguished from news commentary, partisan talk, or fringe discourse. We have suggested predictors to build this critical capability (discussed in 4.1).

It should be noted that the risk of violence against targeted groups is not reduced (and may be increased) when advocates are powerful voices speaking in mainstream contexts. However, where dehumanization is normative and mainstream in a regional context because it is espoused by mainstream politicians or state offices, other forms of politically- and psychologically-informed interventions or challenges may be more effective than content removal.

In our approach, such mainstream groups and content would be placed in the violent extremist category by DMET when regional norms are not considered. All the authors condemn dehumanization against any target in any context. Some authors involved in this report believe platforms could choose to downweigh such groups or content to fringe or partisan on the grounds of regional norms by using exemption functions. For example, dehumanizing homophobia, anti-Semitism, or Islamophobic dialogue advocated by mainstream actors (church leaders, politicians) might be reclassified as partisan or mainstream in certain contexts, when transparently and accountably locally normative. While the Australian Muslim Advocacy Network (AMAN) supports transparency for why certain groups or content is downweighed, it believes that downweighing of dehumanization should be avoided in any regional context by platforms to uphold the overarching obligation under international law not to contribute to the incitement of genocide.

An example of speech that would potentially trigger a violent extremist classification in the absence of regional norm adjustments is provided by political debates over introducing the death penalty for homosexuality in Uganda.[107] For example, the Ugandan Minister for Ethics and Integrity, Simon Lokodo, remarked “Homosexuality is not natural to Ugandans, but there has been a massive recruitment by gay people in schools, and especially among the youth… We want it made clear that anyone who is even involved in promotion and recruitment [of homosexuality] has to be criminalized. Those that do grave acts will be given the death sentence.” Platform providers could consider regional norms despite content being flagged through DMET by transparently exempting state actors from content moderation as discussed in section 5.2 below.

[107] “Uganda Plans to Introduce Death Penalty for Homosexuality with ‘Kill the Gays’ Law,” ABC News, link

Application of DMET

In the following, we illustrate the applicability of DMET by deriving potential instantiations of the cues above for the different levels of ideological engagement together with specific examples for each level. The goal is to clarify the distinction between levels while acknowledging that further efforts are necessary to identify an exhaustive set of instantiations of cues, determine cue up/downweighing methods, sharpen cut-off criteria between levels, and develop strategies to deal with issues such as the niche radicalization of splinter groups. Subsequently, we provide the integrative sample classification of organizations with various degrees of ideological engagement obtained in consultation with global experts.

Illustrative Application of DMET

For partisanship, there would be considerable noise across contexts in how partisan contestation is expressed. At a cognitive level, our starting basket of indicators for partisanship would include simple markers of identification (e.g., use of "we," "us"), us-them distinctions (e.g., "reject," "oppose"), in-group positivity (e.g., "we are good," "we are right"), and out-group negativity (e.g., "they are wrong," "they are bad"). Particular stereotypes that are contextually relevant might be either identified via machine learning or input as cues to screen for (e.g., "Mexican gangs"). While satire and indirection create ambiguity in recognition of cues, specific contextually relevant elements could be coded (e.g., "African gangs"), alongside behavioral indicators of support for in-group actions (e.g., "donate," "volunteer") as well as politicized actions (e.g., "vote," "rally") and artistic contestation (e.g., "protest song," "protest poem"). Signals that indicate partisan group dynamics could comprise moralized grievances (e.g., "justice," "righteousness"), need for significance (e.g., "respect," "be counted") as well as victimization and crisis (e.g., "victim," "crisis," and "threat"); each could constitute initial indicators supporting categorization at this level.

Fringe groups' linguistic and image markers and beliefs would often vary contextually and require local training, but abstract indicators could include cues of dogmatism (e.g., "always," "never"), as well as moral absolutes (e.g., "hero," "villain," "traitor," and "martyr"). Behavioral indicators could include specific contextually relevant insult patterns, narratives, or more abstract categories of coding such as high-arousal negative emotions associated with the out-group-oriented, such as anger, contempt, and disgust.[108] In general, the group dynamics will not likely be transparent to content categorization, although some themes (such as purity and domination) may be available for linguistic coding. In other cases, particular sources or groups could be coded as possessing fringe-characterizing dynamics by experts, and then markers of the source group membership (e.g., jargon and group affiliation terms) could be used to identify content from the fringe

[108] Heerdink, Koning, Doorn, and Van Kleef, "Emotions as Guardians of Group Norms."

actors.

Violent extremists would also engage in dehumanization forms either directly in their language or through the general discourse and conceptions (as elaborated above).

In the second half of 2020, AMAN completed a study of five actors producing significant amounts of blog or pseudo-news content that triggered explicitly dehumanizing and violent responses by users on Facebook and Twitter. That study identified the following markers that were common to all five actors' information operations:

1. Dehumanizing conceptions or conspiracy theories on the actor's website (where applicable) in relation to an identified group ("the out-group") on the basis of a protected characteristic;
2. Repeated features of the headlines and images that are curated for a specific audience, including:
   - Essentializing the target identity through implicating a wide net of identities connected to the protected group (e.g., "Niqab-clad Muslima," "boat migrants," "Muslim professor," "Muslim leader," "Iran-backed jihadis," "Ilhan Omar," "Muslim father");
   - High degree of hostile verbs or actions (e.g., stabs, sets fire) attributed to those subjects;
   - Significant proportion of actor's material acting as "factual proofs" to dehumanizing conceptions about out-group;
   - Potential use of explicitly dehumanizing descriptive language (e.g., frothing-at-the-mouth) or coded extremist movement language with dehumanizing meaning (e.g., invader, a term used in RWE propaganda to refer to Muslims as a mechanically inhuman and barbaric force). However, for the most successful actors, dehumanizing slurs were avoided to maintain legitimacy and avoid detection; and
   - Where there was no dehumanizing language, there was a presence of "baiting" through rhetorical techniques like irony to provoke in-group reactions; and
3. Evidence in the user comment threads of a pattern of hate speech against the out-group.

Markers like these above could be used to train algorithms to identify an information operation intended to dehumanize an out-group over time. Further, GIFCT would be able to compile a list of protected characteristics recognized commonly by member platforms or the United Nations Strategy and Plan of Action on Hate Speech.

Violent extremists could also use images and linguistic markers of out-group violence towards the in-group and contextually relevant images and language of out-group self-defense. Extremists develop narratives legitimizing violence,[109] often by framing the out-group as an enemy who is violent towards them. Hate speech and glorification of violent acts would both be indicators of this level of engagement, as we have noted above.[110] Next to reinforcing a rigid dichotomy that demands that people choose between the forces of good or evil (e.g., ISIS demands that all Sunni Muslims choose to fight with them or against them), they also use apocalyptic linguistic markers to trigger "awakening" in readers.[111] Similarly, terrorists would express beliefs about the legitimacy of killing and the glory of risking sacrificial death. Concrete incitement to violence and physically violent acts provides defining behavioral indicators of the terrorist level in DMET.

Sample Feedback on DMET Classification of Ideologically Engaged Actors

In order to explore the applicability and feasibility of DMET, we reached out to a network of over 20 extremism researchers and counter-extremism advocates through authors' contacts. Our contacts highlighted several aspects of the framework for consideration. They highlighted the simultaneous prevalence of cues from multiple levels and the diffuse nature of some entities as movements rather than groups (e.g., Evangelical Christians). Co-occurrences were most prominent between Level 2 (Violent Extremism) and Level 3 (Terrorism) (e.g., Proud Boys, KKK). The discussion of divergent attributes and diffuse movements particularly appeared for QAnon (categorized by experts across Level 0 – 2) and Incels (Level 1 – 2). Regional differences within movements and the large in-group variability of actors, such as objecting to violence or actively engaging in violence, make movements like QAnon difficult to classify unambiguously.

Some contacts therefore pointed towards the importance of greater flexibility in the analysis. For example, they expressed that some groups would possess attributes of multiple categories (e.g., Institute of Public Affairs as borderline fringe, National Socialist Movement as borderline violent extremist). They also emphasized the multi-faceted approach of various groups assigning them to multiple types of ideological engagement (e.g., right-wing and religious: United Patriots Front; left-wing and separatists: Kurdish movements). Hence, we assume that cross-type patterns need to be acknowledged.

[109] Pratt, "Religion and Terrorism"; Webber and Kruglanski, "The Social Psychological Makings of a Terrorist."

[110] Olteanu et al., "The Effect of Extremist Violence on Hateful Speech Online"; Pratt, "Religion and Terrorism."

[111] Matteo Vergani and Ana-Maria Bliuc, "The Evolution of the Isis' Language: A Quantitative Analysis of the Language of the First Year of Dabiq Magazine," SICUREZZA, TERRORISMO E SOCIETÀ 7 (2015).

Levels of Ideological Engagement	Types of Ideological Engagement				
	Right-Wing	Left-Wing	Religious	Separatist	Single-Issue
Terrorism					
Level 3 Terrorism	Boogaloo Bois Ku Klux Klan National Socialist Network The Base	Sendero Luminoso Fuerzas Armadas Revolucionarias de Colombia	Al-Qaeda Islamic State (Daesh) Jamaah Ansharut Daulah Mujahidin Indonesia Timur (East Indonesia Mujahideen)	Euskadi Ta Askatasuna Irish Republican Army (Provisional Irish Republican Army, Ulster Volunteer Force, Ulster Defence Association)	Army of God Earth Liberation Front
Violent Extremism					
Level 2 Violent Extremism	Blood & Honour Combat 18 United Patriots Front (True blue crew, Lads Society) Oath Keepers Proud Boys Jihad Watch	Antifa Ejército de Liberación Nacional Kurdistan Workers' Party	Forum Pembela Islam	Órganos de Resistencia Territorial	Animal Liberation Front Bundy Family
Non-Violent Extremism					
Level 1 Fringe Group	Australia First party Bharatiya Janata Party National Socialist Movement^	Kurdistan Communities Union Democratic Socialists of America	Brigade Manguni Peoples Temple of the Disciples of Christ Westboro Baptist Church	Greater Idaho Movement Texas Nationalist Movement	American Family Association Andha Chile
Non-Extremism					
Level 0 Partisanship	One Nation Partai Keadilan Sejahtera Tea Party Movement Traditionalist Worker Party	Peoples' Democratic Party (Turkey) Sinn Féin	Wahdah Islamiyah	Scottish National Party	Anti-Vaxxers National Abortion Rights Action League No más AFP (No + AFP)

Notes. ^ trending towards increased ideological engagement

Table 2. Tentative DMET Classification of Ideologically Engaged Actors

Narrative Case-Based Review of Level Differences

In order to offer a hands-on illustration of the individual levels and their mutual differences, we offer a narrative review of individual cases of groups in reference to DMET.

Level 3 (Terrorists): The Base

The DMET-based classification would, for example, echo Canada's recent decision that declared The Base as a terrorist organization. It was founded in 2018 as a neo-Nazi, white-supremacist network that describes itself as an "international survivalist and self-defense network" that seeks to train its members for fighting a race war (Counterextremism, 2021). Cognitive cues as disseminated by the group's leadership denounce the public and system as the enemy. For example, a leading member, Rinaldo Nazzaro, distributed the following message on Telegram, April 8, 2021:

> "Republicans and many White Nationalists think they're fighting for the future of America but they've already lost it and there's no hope of taking it back. The System is irreversibly dominated by the enemy...The System *is* the enemy and the enemy *is* the System—They're inherently and inseparably one and the same now. Some do realize this and hope for a spontaneous collapse which unfortunately will never come. The only victory left to be had is breaking away before it's too late."[112]

Regarding behavioral cues, The Base's leadership has called for members to focus on non-attributable actions that destabilize society. The Base has distributed to its members' manuals for lone-wolf terror attacks, bomb-making, counter-surveillance, and guerilla warfare.[113] Similarly, The Base also promotes a group dynamic with dedicated roles to engage in violent actions. For example, one post by Nazzaro on Telegram on December 21, 2020, reads

"By no later than the 90 day-mark, plan to go on the offensive by clearing and holding the nearest town. You will commandeer the town and this will serve as your new base of operations," before telling followers there may come a time where they will need to kill American citizens if their insurgency is challenged.[114]

Level 2 (Violent Extremists): Oath Keepers

The Oath Keepers would fulfill DMET's characteristics of a violent extremist organization. They are a loosely organized collection of anti-government extremists who are part of the broader anti-government "Patriot" movement with a particular focus on recruiting current and former military members, police officers, and firefighters.[115] The Oath Keepers are driven by conspiracy theories and establish a cognitive glue that promotes violence

[112] "The Base," Counterextremism, Counterextremism.com, link

[113] "The Base," Counterextremism.

[114] "The Base," Counterextremism.

[115] "The Oath Keepers," Anti-Defamation League, link

towards the out-group by asking all members to take a pledge to oppose an allegedly tyrannical American government that will use state forces to control U.S. citizens. The pledge is targeted to refuse or disobey governmental orders to, for example, disarm the society or impose martial law.[116] Prominent Oath Keeper members such as the founder Stewart Rhodes engage in dehumanizing behavior when, for example, declaring migrants or families of legal asylum seekers as an "invasion."[117] Members of the Oath Keepers show considerable personal agency in support of their group and seek to protect its members against the out-group threat, for example by providing armed patrols during the protests in Ferguson or as armed security during land disputes. Similarly, they coalesce with the Constitutional Sheriff and Peace Officers Association (CSPOA) that disputes the federal government's authority and promotes the notion that local sheriffs do not have to obey federal authorities.[118]

Level 1 (Fringe Group): Westboro Baptist Church

DMET would classify the Westboro Baptist Church as a Level 1 Fringe group. The Westboro Baptist Church is an American hyper-Calvinist hate group known for engaging in inflammatory homophobic and anti-American pickets and hate speech against atheists, Jews, Muslims, transgender people, and numerous Christian denominations. The Westboro Baptist Church has an extensive indoctrination system, as evidenced by the comments of a 7-year old member towards an ABC News reporter, saying that those who were destined for eternal damnation included "gays, fags, hundreds and hundreds of Jews."[119] The group has an extensive history of engaging in denigrating antisemitic and anti-gay activities such as over 20,000 respective protests promoting the message that "Any church that allows fags to be members in good standing is a fag church [...] they have created an atmosphere in this world where people believe the lie that God loves everybody."[120] The group holds and enforces strong norms of purity in their beliefs, as most pointedly described by the fact that its founder Fred Waldron Phelps Sr. was excommunicated arguably for diverging from the group's hateful demeanor by suggesting they pursue a kinder approach.[121]

Level 0 (Partisanship): One Nation Group

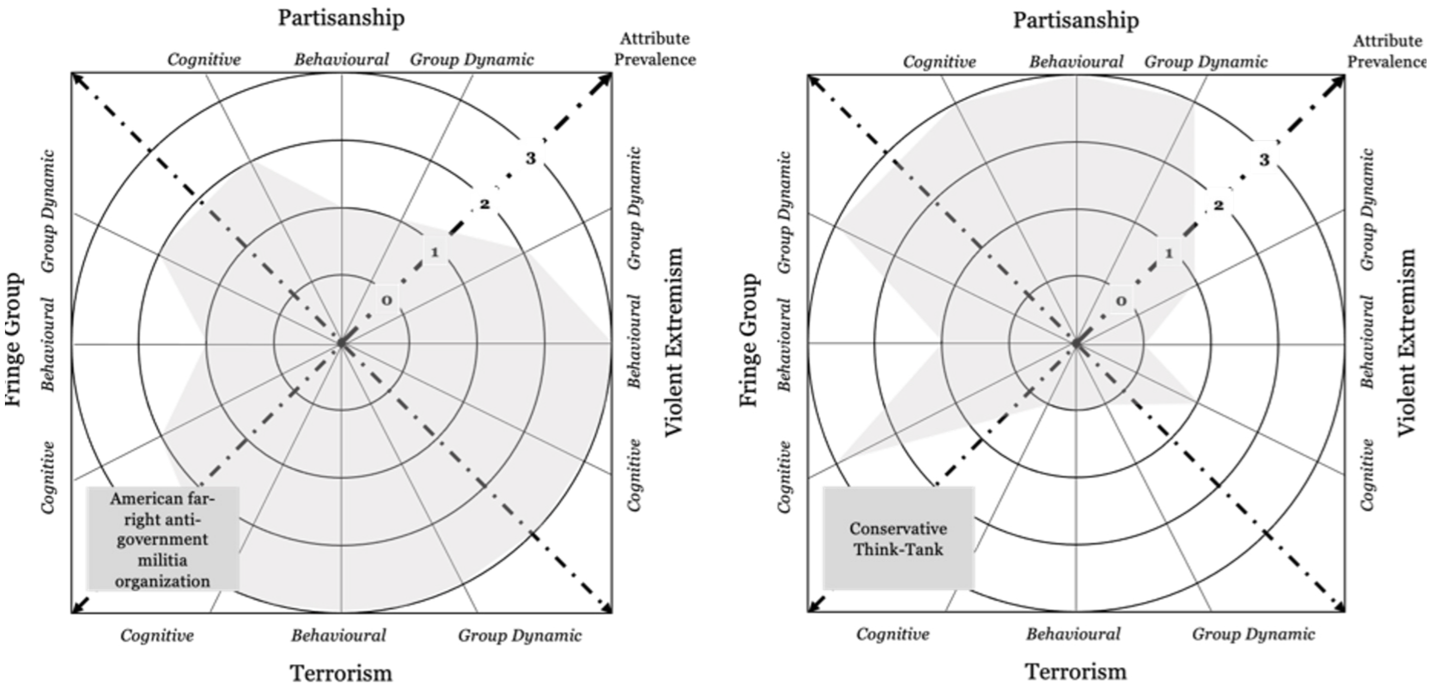
DMET's transition between Level 0 and 1 appears to be more fluid than between other levels. An example of a non-extremist partisan group is the One Nation party (Pauline Hanson's One Nation). It is Australia's far-right political party that was founded in 1997. The founder Pauline Hanson promotes polarizing views of the radical right by using "us-versus-them" language. She holds political grievances that she calls "reverse-racism" or "anti-white" racism and propagates the idea of immigrants and refugees as existential threats to the safety, security, and "culture" of a particular society.[122] Behaviorally, she

[116] "The Oath Keepers," Anti-Defamation League.
[117] John Dougherty, "Oath Keepers 'Call to Action' for Flynn Sentencing a Bust," Southern Poverty Law Center, link.
[118] "The Oath Keepers," Anti-Defamation League.
[119] Glenn Ruppel, Kelsey Myers, and Eamon McNiff, "Raised to Hate: Kids of Westboro Baptist Church," ABC News, link
[120] "Westboro Baptist Church," Anti-Defamation League, link
[121] Victoria Cavaliere, "Founder of Westboro Church in Kansas Excommunicated, on Death Bed - Son," Reuters, link

expresses a strong populist ideology that non-natives must either assimilate and embrace "Australian culture and values" or "go back to where they came from."[123] While PHON has been described as exhibiting hate speech, calls for exclusion, and discrimination, their party obtained 10.27% of the Senate vote in Queensland in the 2019 Federal Election, double its performance nationwide. Its more mainstream acceptance or smaller deviation from the norm in Queensland may warrant its location within the level of partisanship. However, this organization would be considered fringe according to its national levels of political support (Level 1). Australia's ABC News reported that a former PHON candidate later attempted to join The Base out of frustration with the democratic system.[124]

Illustrative Case-Based Empirical Analysis of Level Differences

Beyond the narrative review of individual cases, DMET can potentially be implemented to operationalize and systematically assess cognitive, behavioral, and group dynamic cues of movements, groups, individual actors, or content (e.g., on social media). We envision, for example, assessing social media content regarding their cognitive, behavioral, and group dynamic cues, which can then be aggregated for a particular actor (e.g., individual, group, movement). These empirical analyses can be used to quantify the profile patterns across the different levels of ideological engagements for the following purposes:



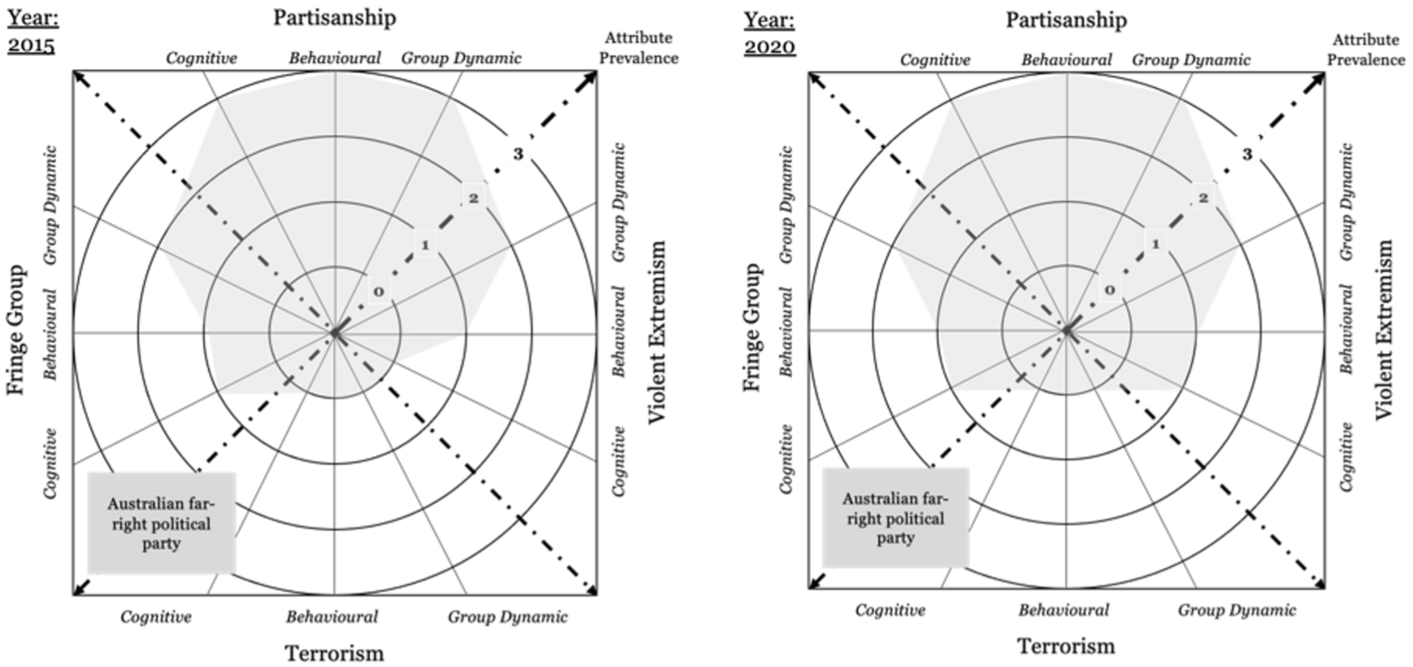
Figures 3 and 4. Conceptual actor profile-specific analyses of the prevalence of definitory cues.

[122] Kurt Sengul, "Pauline Hanson Built a Political Career on White Victimhood and Brought Far-Right Rhetoric to the Mainstream," The Conversation, June 22, 2020, link

[123] "Transcript: Pauline Hanson's 2016 Maiden Speech to the Senate," ABC News, link

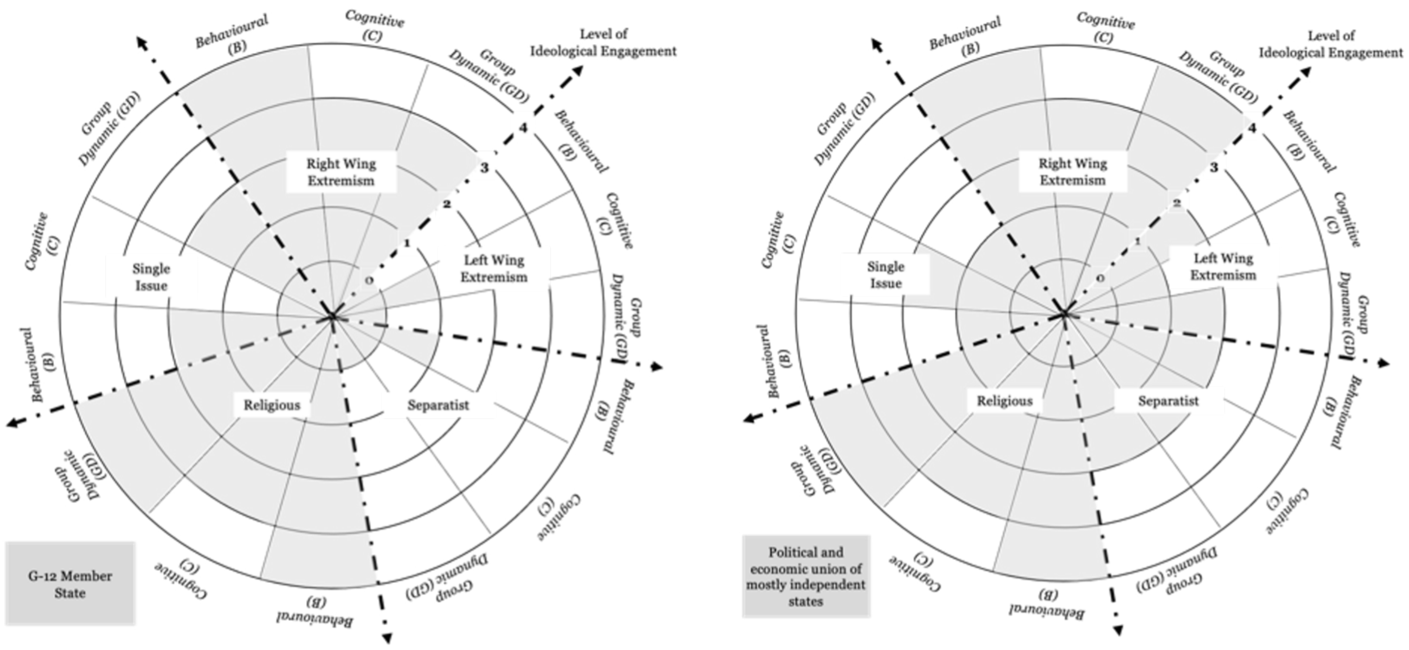
[124] Alex Mann and Kevin Nguyen, "The Base Tapes," ABC News, link

Level of engagement estimation: Given the probabilistic nature of the level defining cues, in-depth DMET-based analyses could assess the prevalence of the proposed attributes per level for different actors (Figures 3 & 4). This would help identify the potential or occurrence of splinter groups. These graphs would highlight whether an organization either solely engages, for example, in fringe activities, or whether others (under the same group name) are already engaged in terrorist activity. Decision makers can transparently assess the profile of different groups or use it to determine the potential threat (parts of) a particular group pose.



Figures 5 and 6. Conceptual actor profile tracking of definitory cues over time.

Time-dependent tracking: DMET can also be used to track and visualize changes over time (Figures 5 & 6). By operationalizing lower levels of non-violent ideological engagement (i.e., Level 0 or 1), DMET enables monitoring relevant ideologically engaged actors before they turn violent (i.e., Level 2 or 3). This could help to transparently flag suspicious actors to decision makers, issue warnings towards respective actors, and systematically re-evaluate the actor profile for de-platforming or delisting decisions.



Figures 6 and 7. Conceptual holistic profile analyses of definitory cues across ideological types.

Holistic profile assessment: Furthermore, DMET can be used to assess all DMET measures for a particular entity holistically. When applied to a particular actor (i.e., individual, group, movement) this can illustrate engagement across different ideological types (e.g., religious fundamentalistic right-wing actors). It can also be applied to create a more general overview of the ideological engagement in the sense of a political barometer in a particular region (Figure 6) and across regions (e.g., state unions, Figure 7).

Discussion

Strategic Application Considerations of DMET

We propose DMET to be used as a foundational framework for classifying forms of extremism and associated extremist content with value beyond approval/removal decisions. We perceive DMET's vital strength in this regard to be in its adaptability to suit varying, fluctuating, or transforming perspectives on what constitutes extremisms. The value-free modularity of DMET remains unbound from fluctuating political, cultural, and/or legal perspectives and provides a depolarized snapshot of attributes on a continuum of extremist magnitudes transparent to platforms, scholars, and the broader public.

To achieve this broad perspective on the existing magnitudes of extremism, considering developments over time and geography requires comprehensive data access. This includes existing historical datasets and consultation with diverse experts and practitioners from academia and industry. This will likely be a modular, iterative approach that assures the implementation feasibility of DMET throughout. For instance, historic and geographic dimensions of extremism might not be included until later stages of the DMET development without impairment of DMET's current operability.

The modularity of DMET that accounts for different understandings and magnitudes of extremism is advantageous for its applicability into content moderation use cases. Capturing different magnitudes of extremism in a multidimensional matrix aids transparency and acts as a decision support system, especially where violent extremists share non-extremist content or vice versa. The source of the non-violent content is identified in relation to the sharing entity (i.e., a particular content's source being identified as VE Actor), aiding in making informed decisions on removing or approving content that would go unnoticed without considering these multiple dimensions of content.

The DMET proposal is qualified by the need for transparency, both prospective (transparency of design) and retrospective (transparency through inspection and explanation), and accountability (managerial and external) to secure public trust.[125] Safeguards should ensure data used to train and develop algorithms is high quality, open to academic scrutiny (including from an AI racial discrimination perspective), and continuously reviewed, corrected, and improved. Normal safeguards against algorithm discrimination in predictive policing (i.e. eliminating variables such as race and religion, or proxies for these variables) are not necessarily going to be appropriate in this context. Although some of the cues relate to cognitions, behaviors, and group dynamics that apply across the ideological spectrum, the data used to train DMET will be drawn from a range of contexts and include racial and religious terms. Checking how representative the

[125] Heike Felzmann et al., "Transparency You Can Trust: Transparency Requirements for Artificial Intelligence between Legal Norms and Contextual Concerns," Big Data & Society 6, no. 1 (2019).

categorization of extremists is in relation to other indicators of prevalence may provide an indicator of bias or generalizability to guide revision.

The DMET should aim to respond to published good practice,[126] principles on artificial intelligence under the Organization for Economic Co-operation and Development (OECD) that were adopted on May 22, 2019, and obligations set out in the European Commission's proposed legislation (e.g., Proposal for a Regulation of the European Parliament and of the council laying down harmonized rules on artificial intelligence, April 21, 2021), as artificial intelligence models to detect terroristic and violent extremist content will be regarded as "high-risk." Mistaken classifications promoted by GIFCT can affect users' rights to free expression on several platforms at once and can even stifle efforts to highlight human rights abuse.[127] Transparent and fair review processes, facilitated by GIFCT to quickly respond to unintended consequences, are also important. It would make it easier for human rights organizations to complain to multiple platforms at once.

We also acknowledge that violent extremist and terrorist labels are highly political and can require platforms to make exemptions for particular actors or content being flagged through DMET. In this approach, as highlighted above, platform providers can toggle content associated with particular groups (e.g., those flagged on U.N. lists) to be always categorized as terrorists, whereas other content (e.g., advocacy for violence by state actors) is not. Similarly, the collective right to self-determination is enshrined in human rights law (see Article 1.1 ICCPR as well as Article 1.1 of ICESCR), and some self-determination movements use violence. If that violence is deemed to be used in "armed conflict" by the International Committee of the Red Cross (ICRC), then humanitarian laws of war apply, and this is not treated as extremism or terrorism. However, some non-state actors will not be listed by the ICRC and governed by humanitarian law because they are engaged in a conflict that does not meet threshold tests.[128] For example, some violent protests will be flagged in DMET, including protests where protestors use violence in response to (or to resist) state violence, including physical violence and life-threatening structural violence. To contend with this gap, platforms could continue to have the discretion to exempt further non-state actors based on exercising self-defense,[129] considering principles such as self-determination, duress, necessity, proportionality, or on the balance of other fundamental human rights. Our hope would be that if a group's content is flagged in DMET for the use of dehumanization or advocacy of violence but exempted by platforms, they would be transparent about exemptions made and provide reasons.

Making a decision based on the balance of fundamental human rights may be required to provide an enduring mechanism for managing conflict between differing world views and

[126] Felzmann et al., "Transparency You Can Trust."
[127] Abdul Rahman Al Jaloud et al., "Caught the Net: The Impact of "Extremist" Speech Regulations on Human Rights Content," Electronic Frontier Foundation ed. JIllian C. York (2019), link
[128] See the categorization of armed conflict as proposed by UNODC: link
[129] Ben Saul, "Defending 'Terrorism': Justifications and Excuses for Terrorism in International Criminal Law," Australian Yearbook of International Law 25 (2008).

claims. For example, groups that wish to express themselves, their beliefs, and exercise their fundamental rights (such as the right to parent their children and choose schooling according to their beliefs) are protected by human rights law to do so. Groups are not protected to infringe upon the fundamental rights of others, and such behavior would begin to animate DMET cues.

Noting that DMET focuses on deviation from social norms, it is important to consider that a violent protest will deviate less from mainstream social norms in some regional contexts. For example, where mass popular protest movements feature violent elements and advocacy of violence against law enforcement and the state, the scale of people involved will mean that their behavior may not be flagged as non-normative, extreme, or radical by regional standards.

Technical Implementation Considerations of DMET

In our approach, large baskets of indicators would be associated probabilistically with each level (e.g., cognitive stereotypes, dehumanizing language, calls for violence) and will be used to develop models that algorithmically classify groups (or content) into categories, with each cue weighed according to its ability to discriminate in particular contexts defined by the other cues as well as input from the platform provider where desired. We imagine an incoming stream of content coded for the indicator cues and the groups involved via machine learning in a process that would be more error-prone at first and be refined over time and regionally to produce context-specific accuracy, which would in turn decay as diagnostic attributes changed over time until relearned dynamically. Especially in the beginning, this will require extensive human oversight, for example in order to deal with expected inaccuracies of automated machine learning algorithms when dealing with linguistic markers for irony, sarcasm, or subtle dehumanization.

Content and groups would be classified probabilistically into categories where cues have established sufficient discriminant validity and high confidence (e.g., with explicit calls to violence, or when the content is sourced from a group identified as a terrorist organization, or as a state actor or journalistic or academic source). More commonly, groups and individuals would be classified based on profiles established via multiple content posts with increasing confidence over time, with each content item or group reciprocally associated with transparent certainty/uncertainty scores based on a profile of attributes, which could be available to platforms as an output.

As a next step, DMET could train a machine learning algorithm that identifies the different level cues in online content. Due to the linguistic challenges and subjective interpretation of the content in this unique context, unsupervised learning approaches are likely to provide misleading results. There is a need for more sophisticated models, and building such models requires generating a labeled training data set from scratch. A potential cold start problem of insufficient data for initially training the algorithm could be overcome by collecting social media content from the groups, building on those identified by experts as above. This data could comprise social media posts and comments, memes, or content from external sites (e.g., extremist websites, blogs) introduced into conversations through hyperlinks. The accuracy of the machine learning algorithm will mainly depend on the (1) clarity and consistency of classification rule (the coding scheme), (2) quality and size of labeled data, and (3) finding the proper feature representation.

Subsequently, we could analyze the actor-specific distribution of individual messages across the different levels to establish and identify communication patterns. By applying the previously developed machine learning algorithm to new data, we could expand the available content coded data. This would help perform a ROC analysis to determine extremism cut-off scores between levels. The regression weights for the individual cues could also serve as an indicator for the up-/downweighing of individual cues. By choosing to downweigh or upweigh particular dimensions, platform providers can establish local profiles of tolerance (e.g., no hate speech at all versus this group; versus hate speech tolerated against this group, due to its being normative in this context) in a way that is transparent and able to be accountable or engaged with dialogue. Platform providers may also opt for transparently and accountably in exempting certain actors such as state organizations or religious groups ex officio because their views (e.g., the Ugandan minister above) are mainstream rather than extreme in the regional context.

We could complement the analysis by using metadata that contains information on the connection between entities to consider the hierarchical structure between individual extremists (potentially) embedded in extremist groups who are themselves nested in ideological movements.

Benefits of DMET

We understand extremism as a dimensional concept, with terrorism as the most deviant pole from the regional norm. DMET supports:

- **Ideological fairness:** equal opportunity for all entities to be classified as terrorist based on the generally applicable cognitive, behavioral, and group dynamic cues;
- **Global applicability and scalability:** definitory cues transcend geographical, cultural, and political borders and can be applied relative to relevant reference norms;
- **Update frequency:** Observable changes in cognitions, behaviors, or group dynamics can be captured through near real-time updates;
- **Transparency:** Classifying actors according to DMET categories based on their degree of ideological engagement enables transparency and accountability in regulatory decisions;
- **Surfacing states' role and influence:** Current definitions of extremists or terrorists

often exclude state actors. DMET potentially classifies any kind of actor without consideration of their societal role. Platforms can navigate these situations by making transparent exemptions in reference to the classifications proposed by DMET to justify their decision making;

- **Reduced probability of misclassification errors for (non-) violent extremisms:** A more nuanced understanding of the degrees of ideological engagement and the potential sub-groups reduces the probability of wrong decisions (classifying regular users/content as terrorist or failing to identify terrorists as such); and
- **Attention to violence in all its forms:** Many existing legal frameworks are so piecemeal or narrow that they deprioritize and overlook the experience of victims and communities targeted by terroristic and violent extremist violence. DMET contemplates the full continuum of violence that occurs in the violent denial of diversity, including structural and psychological violence. Importantly, it recognizes serial or systematic dehumanization of an out-group as an attribute of violent extremism.

Subsequently, DMET addresses contemporary challenges of extremism classification and associated content moderation approaches, including the lack of consent on universal definitions of extremism, bias, and deficient objectivity on different magnitudes of extremism.[130] DMET's multidimensional approach enables the aggregation of various lists and dimensions to allow biases in views of extremism (e.g., exemptions for certain actors) to be more transparent and accountable. Moreover, DMET unlocks the possibility of a 360-degree context view of extremism irrespective of the limitations of individual extremism lists and allows for tracking the development in terms of (de-)radicalization over time through continuous assessments among the spectrum of ideological engagement.

Boundary Conditions of DMET

- Dimensionality of attributes: DMET ascribes attributes to particular levels to capture different degrees of severity. It needs to be acknowledged that these attributes themselves can also be dimensional (e.g., expressing blame for negative events can be more or less rampant). Similarly, actors might, for example, dehumanize a group by using a multitude of cues that holistically dehumanize the target without making it explicit in one singular instance. The dimensionality of attributes needs to be empirically assessed to be statistically considered through measures of item difficulty and item discrimination. The individually classified instances then need to be holistically considered for each entity. This would also enable DMET classifications to record an actor's tendency of either trending towards a higher (or lower) DMET level of ideological

[130] Meserole and Byman, "Terrorist Definitions and Designations Lists."

engagement or of being stable within the level.

- Probability of attributes: Similarly, DMET ascribes attributes to levels of ideological engagement where we expect their highest probability of prevalence. We acknowledge that ideologically engaged actors (i.e., individuals, groups, movements) can simultaneously demonstrate cognitive, behavioral, and group dynamic cues from multiple levels or not express particular cues from the level at which they are classified. The classification of ideological actors according to DMET requires an empirical assessment of the expectable probabilities of cues per level and their respective level-determinant weight (i.e., up- or downweighing of attributes).
- Combinability of types: While DMET distinguishes five common types of ideological engagement, we understand that actors (i.e., individuals, groups, movements) can simultaneously follow different types of ideologies (e.g., nationalism in combination with religious fundamentalism). Hence, DMET classifications need to acknowledge the expressivity of characteristics across multiple ideological types per actor.
- Fragmentation of actors: Different actor organizations (i.e., groups, movements) can include splinter groups or individuals that diverge from the characteristics of the overall organization (e.g., enact or support violent behavior as opposed to the general movement). These individuals or groups can either emerge as splinter groups or lone-wolf actors alongside the general movement or relative to their location (e.g., violent in one country, non-violent in another). Classifications according to DMET need to acknowledge the relatedness of the individual groups or actors to the higher-level organization (e.g., via metadata) while considering their individual particularities.
- Time and cultural specificity: Actors (i.e., individuals, groups, movements) as well as the expression of cues evolve. Actors might become more or less ideologically engaged, splinter-off into different sub-groups, or form coalitions with other movements. Similarly, how cues are expressed can change as words can adopt different meanings over time, as the societal acceptability of terms evolves, or as terms have regionally specific meanings (e.g., reference to Odin in Nordic nationalist groups versus the rest of the world). DMET classifications need to be considered at particular points in time, in particular regions, and regularly re-evaluated in predetermined time intervals (e.g., to inform de-platforming or readmittance and delisting decisions).
- Complexity of societal norms: DMET characteristics need to be assessed against relevant societal norms, which is intended to support its general applicability. However, regional norms may vary substantially.

Conclusion

This paper has put forward a proposal for extending the binary understanding of terrorism (versus non-terrorism) with a Dynamic Matrix of Extremisms and Terrorism (DMET) that builds upon the notion of an underlying continuum of ideological engagement to address the intersection of extremism types (and associated extremist content) that lead to ineffective content tagging. DMET considers different types of ideological engagement and different levels, identified using cognitive and behavioral attributes and attributes of group dynamics. DMET is dynamic as it can be adapted to accommodate region- and time-specific notions of ideological engagement. The goal of DMET is to enable platform providers to make transparent and accountable decisions about engaging with content and groups so that violent extremist and terrorist content can be identified in a way that makes explicit the criteria and dimensions underlying the categorization and allows areas of contestation and change to be identified.

DMET Graphics and Visualizations

Reference

ABC News. "Transcript: Pauline Hanson's 2016 Maiden Speech to the Senate." ABC News, link

———. "Uganda Plans to Introduce Death Penalty for Homosexuality with 'Kill the Gays' Law." ABC News, link

Abrams, Dominic, José M Marques, Nicola Bown, and Michelle Henson. "Pro-Norm and Anti-Norm Deviance within and between Groups." Journal of personality and social psychology 78, no. 5 (2000): 906.

Abrams, Dominic, Giovanni A Travaglino, José M Marques, Ben Davies, and Georgina Randsley de Moura. "Collective Deviance: Scaling up Subjective Group Dynamics to Superordinate Categories Reveals a Deviant Ingroup Protection Effect." Journal of Personality and Social Psychology (2021).

Ackerman, Gary A, Jun Zhuang, and Sitara Weerasuriya. "Cross-Milieu Terrorist Collaboration: Using Game Theory to Assess the Risk of a Novel Threat." Risk analysis 37, no. 2 (2017): 342-71.

ADL. "The Oath Keepers " Anti Defamation League, link

———. "Westboro Baptist Church." Anti-Defamation League, link

Aytaç, S Erdem, Ali Çarkoğlu, and Ezgi Elçi. "Partisanship, Elite Messages, and Support for Populism in Power." European Political Science Review 13, no. 1 (2021): 23-39.

Bakker, Bert N., Yphtach Lelkes, and Ariel Malka. "Understanding Partisan Cue Receptivity: Tests of Predictions from the Bounded Rationality and Expressive Utility Perspectives." The Journal of Politics 82, no. 3 (2020): 1061-77.

Bandura, Albert. "Moral Disengagement in the Perpetration of Inhumanities." Personality and Social Psychology Review 3, no. 3 (1999/08/01 1999): 193-209.

Baumeister, Roy F. Evil: Inside Human Cruelty and Violence. WH Freeman/Times Books/Henry Holt & Co, 1996.

Berger, John M. "Deconstruction of Identity Concepts in Islamic State Propaganda: A Linkage-Based Approach to Counter-Terrorism Strategic Communications." 1-21. The Hague, Netherlands: EUROPOL, 2017.

Borum, Randy. "Radicalization into Violent Extremism I: A Review of Social Science Theories." Journal of strategic security 4, no. 4 (2011): 7-36.

Cavaliere, Victoria. "Founder of Westboro Church in Kansas Excommunicated, on Death Bed - Son." Reuters, link

Christmann, Kris. "Preventing Religious Radicalisation and Violent Extremism: A Systematic Review of the Research Evidence." (2012).

Cook, David. "The Implications of "Martyrdom Operations" for Contemporary Islam." The Journal of Religious Ethics 32, no. 1 (2004): 129-51.

Coulson, Andrew. "Education and Indoctrination in the Muslim World." Policy Analysis 29 (2004): 1-36.

Counterextremism. "The Base." Counterextremism.com, link

Dougherty, John. "Oath Keepers 'Call to Action' for Flynn Sentencing a Bust." Southern Povery Law Center, link

Feddes, Allard R., Lars Nickolson, Liesbeth Mann, and Bertjan Doosje. Psychological Perspectives of Radicalization. London: Routledge, 2020.

Felzmann, Heike, Eduard Fosch Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. "Transparency You Can Trust: Transparency Requirements for Artificial Intelligence between Legal Norms and Contextual Concerns." Big Data & Society 6, no. 1 (2019): 2053951719860542.

Giner-Sorolla, Roger, Bernhard Leidner, and Emanuele Castano. "Dehumanization, Demonization, and Morality Shifting: Paths to Moral Certainty in Extremist Violence." Extremism and the psychology of uncertainty (2012): 165-82.

Goffman, Erving. Stigma: Notes on the Management of Spoiled Identity. Simon and Schuster, 2009.

Graham, Roderick. "Inter-Ideological Mingling: White Extremist Ideology Entering the Mainstream on Twitter." Sociological Spectrum 36, no. 1 (2016/01/02 2016): 24-36.

Haslam, Nick. "Dehumanization: An Integrative Review." Personality and Social Psychology Review 10, no. 3 (2006): 252-64.

Heerdink, Marc W, Lukas F Koning, Evert A Van Doorn, and Gerben A Van Kleef. "Emotions as Guardians of Group Norms: Expressions of Anger and Disgust Drive Inferences About Autonomy and Purity Violations." Cognition and emotion 33, no. 3 (2019): 563-78.

Hogg, M. A., and S. A. Reid. "Social Identity, Self-Categorization, and the Communication of Group Norms." [In English]. Communication Theory 16, no. 1 (Feb 2006): 7-30.

Hogg, M. A., and D. J. Terry. "Social Identity and Self-Categorization Processes in Organizational Contexts." [In English]. Academy of Management Review 25, no. 1 (Jan 2000): 121-40.

Hogg, Michael A, Arie Kruglanski, and Kees Van den Bos. "Uncertainty and the Roots of Extremism." Journal of Social Issues 69, no. 3 (2013): 407-18.

Holbrook, Donald. “Designing and Applying an ‘Extremist Media Index’.” Perspectives on Terrorism 9, no. 5 (2015): 57-68.

———. “Far Right and Islamist Extremist Discourses: Shifting Patterns of Enmity.” Extreme Right Wing Political Violence and Terrorism (2013): 215-37.

Horgan, John G, Max Taylor, Mia Bloom, and Charlie Winter. “From Cubs to Lions: A Six Stage Model of Child Socialization into the Islamic State.” Studies in Conflict & Terrorism 40, no. 7 (2017): 645-64.

Jaloud, Abdul Rahman Al, Hadi Al Khatib, Jeff Deutch, Dia Kayyali, and Jillian C. York. “Caught the Net: The Impact of “Extremist” Speech Regulations on Human Rights Content.” edited by JIllian C. York, 2019.

Jegede, Ayodele Samuel. “What Led to the Nigerian Boycott of the Polio Vaccination Campaign?” [In eng]. PLoS medicine 4, no. 3 (2007): e73-e73.

Keatinge, Tom, and Florence Keen. “Social Media and (Counter) Terrorist Finance: A Fund-Raising and Disruption Tool.” Studies in Conflict & Terrorism 42, no. 1-2 (2019/02/01 2019): 178-205.

Keatinge, Tom, Florence Keen, and Kayla Izenman. “Fundraising for Right-Wing Extremist Movements.” The RU.S.I Journal 164, no. 2 (2019/02/23 2019): 10-23.

Kinder, Donald R, and D Roderick Kiewiet. “Economic Discontent and Political Behavior: The Role of Personal Grievances and Collective Economic Judgments in Congressional Voting.” American Journal of Political Science (1979): 495-527.

Knobloch-Westerwick, Silvia, and Simon M Lavis. “Selecting Serious or Satirical, Supporting or Stirring News? Selective Exposure to Partisan Versus Mockery News Online Videos.” Journal of Communication 67, no. 1 (2017): 54-81.

Koehler, Daniel. “Radical Groups’ Social Pressure Towards Defectors: The Case of Right-Wing Extremist Groups.” Perspectives on Terrorism 9, no. 6 (2015): 36-50.

Kruglanski, Arie W, Katarzyna Jasko, Marina Chernikova, Michelle Dugas, and David Webber. “To the Fringe and Back: Violent Extremism and the Psychology of Deviance.” American Psychologist 72, no. 3 (2017): 217.

Kruglanski, Arie W., Michele J. Gelfand, Jocelyn J. Bélanger, Anna Sheveland, Malkanthi Hetiarachchi, and Rohan Gunaratna. “The Psychology of Radicalization and Deradicalization: How Significance Quest Impacts Violent Extremism.” Political Psychology 35 (2014): 69-93.

Lentini, Peter. “The Australian Far-Right: An International Comparison of Fringe and Conventional Politics.” In The Far-Right in Contemporary Australia, edited by Mario Peucker and Debra Smith, 19-51. Singapore: Springer Singapore, 2019.

Lupu, Noam. “Party Polarization and Mass Partisanship: A Comparative Perspective.” Political Behavior 37, no. 2 (2015/06/01 2015): 331-56.

Mandel, David R. “The Role of Instigators in Radicalization to Violent Extremism.” Psychosocial, Organizational, and Cultural Aspects of Terrorism: Final Report to NATO HFM140/RTO. Brussels: NATO (2011).

Mann, Alex, and Kevin Nguyen. “The Base Tapes.” ABC News, link

Marczak, Nikki. “A Century Apart: The Genocidal Enslavement of Armenian and Yazidi Women.” In A Gendered Lens for Genocide Prevention, edited by Mary Michele Connellan and Christiane Fröhlich, 133-62. London: Palgrave Macmillan UK, 2018.

Maynard, Jonathan Leader, and Susan Benesch. “Dangerous Speech and Dangerous Ideology: An Integrated Model

for Monitoring and Prevention .” Genocide Studies and Prevention: An International Journal 9, no. 3 (2016): 70-95.

McCauley, Clark, and Sophia Moskalenko. “Mechanisms of Political Radicalization: Pathways toward Terrorism.” Terrorism and political violence 20, no. 3 (2008): 415-33.

Meserole, Chris, and Daniel L. Byman. “Terrorist Definitions and Designations Lists What Technology Companies Need to Know.” In Global Research Network on Terrorism and Technology, 2019.

Moghaddam, F. M. “The Staircase to Terrorism a Psychological Exploration.” American Psychologist 60, no. 2 (2005): 161-69.

Morrison, Sara. “Westboro Baptist Church’s Founder Is Dying, Excommunicated.” The Atlantic, link

Neumann, Peter R. “The Trouble with Radicalization.” International affairs 89, no. 4 (2013): 873-93.

Oegema, Dirk, and Bert Klandermans. “Why Social Movement Sympathizers Don’t Participate: Erosion and Nonconversion of Support.” American Sociological Review (1994): 703-22.

Olteanu, Alexandra, Carlos Castillo, Jeremy Boy, and Kush Varshney. “The Effect of Extremist Violence on Hateful Speech Online.” Paper presented at the Proceedings of the International AAAI Conference on Web and Social Media, 2018.

Polo, Sara M.T. “The Quality of Terrorist Violence: Explaining the Logic of Terrorist Target Choice.” Journal of peace research 57, no. 2 (2020): 235-50.

Pratt, Douglas. “Religion and Terrorism: Christian Fundamentalism and Extremism.” Terrorism and Political Violence 22, no. 3 (2010): 438-56.

Ruppel, Glenn, Kelsey Myers, and Eamon McNiff. “Raised to Hate: Kids of Westboro Baptist Church.” ABC News, link

Sageman, Marc. Understanding Terror Networks. University of Pennsylvania press, 2011.

Saul, Ben. "Defending 'Terrorism': Justifications and Excuses for Terrorism in International Criminal Law." Australian Yearbook of International Law 25 (2008): 177-226.

Schmid, Alex P. "Violent and Non-Violent Extremism: Two Sides of the Same Coin." International Centre for Counter-Terrorism (ICCT) Research Paper (2014): 1-29.

Sengul, Kurt. "Pauline Hanson Built a Political Career on White Victimhood and Brought Far-Right Rhetoric to the Mainstream." June 22, 2020 2020.

Sidanius, J., C. Van Laar, S. Levin, and S. Sinclair. "Ethnic Enclaves and the Dynamics of Social Identity on the College Campus: The Good, the Bad, and the Ugly." [In English]. Journal of Personality and Social Psychology 87, no. 1 (Jul 2004): 96-110.

Sides, John, Michael Tesler, and Lynn Vavreck. Identity Crisis: The 2016 Presidential Campaign and the Battle for the Meaning of America. Princeton University Press, 2019.

Sternberg, Robert J. "A Duplex Theory of Hate: Development and Application to Terrorism, Massacres, and Genocide." Review of General Psychology 7, no. 3 (2003): 299-328.

Stott, C., P. Hutchison, and J. Drury. "'Hooligans' Abroad? Inter-Group Dynamics, Social Identity and Participation in Collective 'Disorder' at the 1998 World Cup Finals." [In English]. British Journal of Social Psychology 40 (Sep 2001): 359-84.

Tosini, Domenico. "Calculated, Passionate, Pious Extremism: Beyond a Rational Choice Theory of Suicide Terrorism." Asian Journal of Social Science 38, no. 3 (2010): 394-415.

van Zomeren, M., T. Postmes, and R. Spears. "Toward an Integrative Social Identity Model of Collective Action: A Quantitative Research Synthesis of Three Socio-Psychological Perspectives." [In English]. Psychological Bulletin 134, no. 4 (Jul 2008): 504-35.

Vergani, Matteo, and Ana-Maria Bliuc. "The Evolution of the Isis' Language: A Quantitative Analysis of the Language of the First Year of Dabiq Magazine." SICUREZZA, TERRORISMO E SOCIETÀ 7 (01/01 2015).

Verkuyten, M., and A. De Wolf. "The Development of in-Group Favoritism: Between Social Reality and Group Identity." [In English]. Developmental Psychology 43, no. 4 (Jul 2007): 901-11.

Victoroff, Jeff, Janice R. Adelman, and Miriam Matthews. "Psychological Factors Associated with Support for Suicide Bombing in the Muslim Diaspora." Political Psychology 33, no. 6 (2012): 791-809.

Vinciarelli, Alessandro, Maja Pantic, Hervé Bourlard, and Alex Pentland. "Social Signal Processing: State-of-the-Art and Future Perspectives of an Emerging Domain." In Proceedings of the 16th ACM international conference on Multimedia, 1061–70. Vancouver, British Columbia, Canada: Association for Computing Machinery, 2008.

Webber, David, and Arie W Kruglanski. "The Social Psychological Makings of a Terrorist." Current opinion in psychology 19 (2018): 131-34.

West, Emily A, and Shanto Iyengar. "Partisanship as a Social Identity: Implications for Polarization." Political Behavior (2020): 1-32.

Wetts, Rachel, and Robb Willer. "Who Is Called by the Dog Whistle? Experimental Evidence That Racial Resentment and Political Ideology Condition Responses to Racially Encoded Messages." Socius 5 (2019): 2378023119866268.

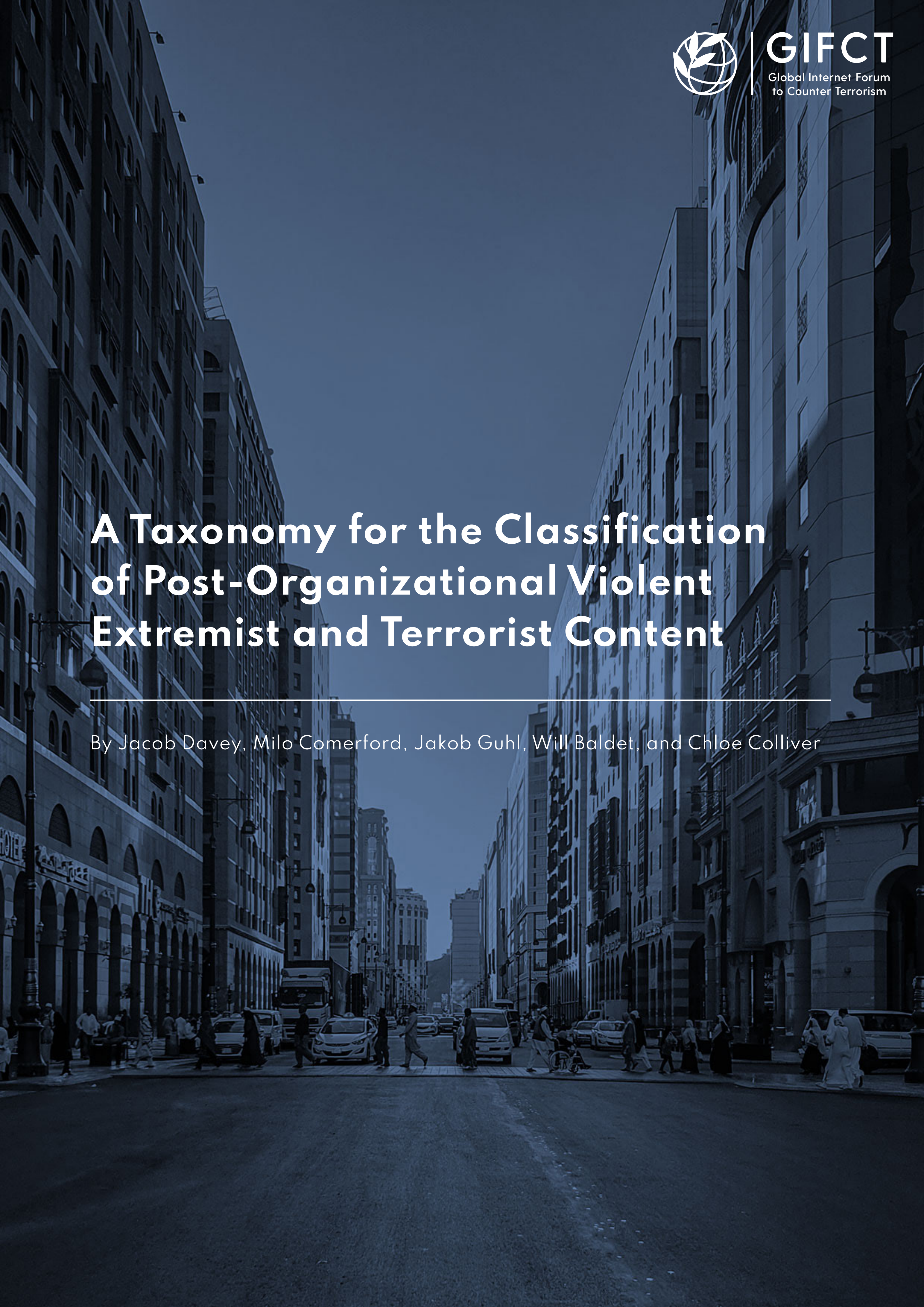
White, Jonathan, and Lea Ypi. The Meaning of Partisanship. Oxford University Press, 2016.

Wibisono, Susilo, Winnifred R Louis, and Jolanda Jetten. "A Multidimensional Analysis of Religious Extremism." Frontiers in psychology 10 (2019): 2560.

Winter, Charlie, Peter Neumann, Alexander Meleagrou-Hitchens, Magnus Ranstorp, Lorenzo Vidino, and Johanna Fürst. "Online Extremism: Research Trends in Internet Activism, Radicalization, and Counter-Strategies." International Journal of Conflict and Violence (IJCV) 14, no. 2 (2020): 1-20.

Wintrobe, Ronald. Rational Extremism: The Political Economy of Radicalism. Cambridge University Press, 2006.

Wright, S. C., D. M. Taylor, and F. M. Moghaddam. "Responding to Membership in a Disadvantaged Group - from Acceptance to Collective Protest." [In English]. Journal of Personality and Social Psychology 58, no. 6 (Jun 1990): 994-1003.



GIFCT
Global Internet Forum
to Counter Terrorism

A Taxonomy for the Classification of Post-Organizational Violent Extremist and Terrorist Content

By Jacob Davey, Milo Comerford, Jakob Guhl, Will Baldet, and Chloe Colliver

A Taxonomy for the Classification of Post-organizational Violent Extremist and Terrorist Content

By Jacob Davey, Milo Comerford, Jakob Guhl, Will Baldet, and Chloe Colliver

Summary

This report outlines a prototype taxonomy for classifying terrorist and violent extremist content. It is designed to inform content moderation decisions made by social media platforms, including adjustments to the hash-sharing database of the Global Internet Forum to Counter Terrorism (GIFCT), which provides unique digital "fingerprints" of known terrorist content which has been removed from social media platforms.

In particular, this taxonomy is developed in recognition of "post-organizational" violent extremism and terrorism where the influence or direction of activity by particular groups or organizations is ambiguous or loose. Accordingly, it is designed to be group-agnostic and is instead shaped around the analysis of influential violent extremism and terrorism content beyond that produced by proscribed terrorist organizations.

The creation of this taxonomy was informed by analysis of content shared in post-organizational violent extremist and terrorist spaces online, and online material which has been referenced in the conviction of terror offenders in the United Kingdom and inquiries into terrorist attacks. This included analysis of the "Terrorgram" network of violent white-supremacist channels on Telegram; the conviction of Jack Reed, the youngest individual to be convicted of terror offences in the U.K.; the Royal Commission of Inquiry into the 2019 Christchurch attack; a cache of online content maintained by supporters of ISIS; and analysis of material referenced in the convictions of Islamist terrorists in the U.K.

Based on assessments of these emblematic case studies relating to contemporary post-organizational terrorism, our group-agnostic taxonomy divides violent extremist and terrorist content into three overarching categories: "inspirational" content designed to reinforce a violent extremist mind-set; "ideological" content designed to further a violent extremist world view; and "instructional" content designed to inform operational aspects of violent extremist activity.

This paper provides an overview of the taxonomy and the process behind its creation, a discussion of the parameters of content included in the content and "edge cases," case studies demonstrating its application, and considerations around its practical implementation.

Background: the challenge of post-organizational violent extremism and terrorism

In recent years terrorism and violent extremism across the ideological spectrum have been marked by a "post-organizational" trend.[131] Membership of and support for particular groups has become more ambiguous, with online activity facilitating the growth of more fluid transnational movements. Attacks are committed by individuals with no, or very loose, connection to specific organizations, and violent extremists instead draw on a shared culture and ideology.

This post-organizational phenomenon is not new. The notions of "leaderless resistance" and "leaderless jihad"[132] were first discussed decades ago by extremist ideologues such as white-supremacist Louis Beam Jr. and the al-Qaeda-linked Abu Musab al-Suri.[133] However, recent high-profile terrorist attacks in New Zealand, the U.S., Germany, and Norway have shone a light on self-radicalizing logistically autonomous individuals with little or no relationship with proscribed terrorist groups, but rather connections to loose transnational extremist networks largely operating online. As Colin Clarke and Bruce Hoffman have noted in the context of U.S. domestic violent extremism, organizational structure is becoming less relevant as "a confluence of ideological affinities is [becoming] more powerful in inspiring and provoking violence than the hierarchical terrorist organizational structures of the past."[134]

Despite the fracturing and franchising of violent extremist movements and the proliferation of decentralized online extremist spaces, responses to terrorist content online are still hampered by rigid organizational conceptions of the challenge.

In particular, post-organizational dynamics strain responses that focus solely on the proscription of specific organizations. Moves have been welcomed to proscribe extreme right-wing groups as terrorist organizations in national contexts, such as National Action in the U.K. and Blood & Honour in Canada, as well as the U.S. listing of its first "Racially and Ethnically Motivated" foreign terrorist organization – the Russian Imperial Movement. But the result has been that groups are banned in some countries but not others, even if – like Combat 18 – they have transnational membership.

Furthermore, when groups appear they are often relatively short-lived, with new movements springing up and drawing inspiration from similar core texts and ideologies. For example, in 2020 the U.K. proscribed the neo-Nazi group Feuerkrieg Division after it

[131] Joe Mullhall, "A Post-Organisational Far Right?," Hope Not Hate, published 2018, link; Milo Comerford, "Confronting the Challenge of 'Post-Organisational' Extremism," Observer Research Foundation, August 19, 2020, link

[132] While al-Suri drove the development of strategic thinking along these lines a salafi-jihadist context, the term 'leaderless jihad' originated in Marc Sageman, Leaderless jihad: Terror networks in the twenty-first century, (Philadelphia: University of Pennsylvania Press, 2011).

[133] JM Berger, "The Strategy of Violent White Supremacy Is Evolving," The Atlantic, August 7, 2019, link

[134] Bruce Hoffman and Colin Clarke, "The Next American Terrorist," The Cipher Brief, July 2, 2020, link

had officially disbanded.[135]

This rapid evolution of groups and movements means that proscription-based approaches follow a "whack-a-mole" dynamic, constantly re-calibrating to address the latest iteration of a movement rather than addressing its roots. Accordingly, relatively slow-moving proscription-based responses to terrorism are not effectively equipped to deal with the current dynamic nature of terrorist mobilization.

While tech companies have been developing their own internal guidelines and terms of service around "hateful" and "dangerous" groups, specific policies around violent extremism and terrorism are partly hamstrung by the limitations of international lists of proscribed terrorist groups, such as the U.N. Designated Terror Groups list, which are focused almost exclusively on ISIS and al-Qaeda related threats. Structurally, international counter-terrorism efforts are therefore still geared towards combating an organized Islamist threat. This has had an indirect effect on the scope and definitional framework lying behind the combined efforts of tech companies through GIFCT and in particular its hash-sharing database.

The hash-sharing database provides unique digital fingerprints, or "hashes," of known violent terrorist imagery or recruitment videos.[136] This tool has increased cross-industry cooperation on the detection and possible removal of illegal terrorist content. GIFCT's July 2020 Transparency Report reveals the sort of content currently covered by the tool.[137] This includes public postings which represent an imminent, credible threat of violence; depictions of graphic violence against defenseless people; glorification of terrorist acts; material that seeks to recruit individuals or give them operational guidance; and content from specific attacks (e.g. those in Christchurch, Halle, and Glendale).

However, such tools have not been designed to tackle the increasingly diffuse "post-organizational" threat emerging from extremism across the ideological spectrum. This briefing will seek to answer the question of how we develop policy frameworks that move beyond a group-centered approach to understanding the threat from violent extremist groups, while ensuring approaches remain robust, transparent, and protective of fundamental freedoms.

[135] Lizzie Dearden, "Why has Britain banned a neo-Nazi terrorist group that 'no longer exists'?," The Independent, July 14, 2020, link
[136] GIFCT, "Joint Tech Innovation," GIFCT, published 2020, link
[137] GIFCT, GIFCT Transparency Report July 2020, GIFCT, July 2020, link

Designing a group-agnostic framework for classifying post-organizational terrorist and violent extremist content

To help inform responses to violent extremism and terrorism online (including, but not limited to hash-sharing approaches), we have devised a theoretical framework for classifying content aimed at moving beyond solely group-centered approaches.

The creation of this framework was derived from analysis of a number of case studies of unequivocally violent extremist environments – all of which demonstrate the real-world fluidity associated with classifying violent extremist content, and the shortcomings of purely group-based analyses.

In an effort to develop an ideologically agnostic framework, ISD analysts examined content shared in several violent extremist spaces online from across the extremist ideological spectrum, as well as content that has been used as evidence in the prosecution of terror offenders or inquiries into terrorist attacks, much of which currently falls outside the scope of GIFCT's existing hash-sharing database.

The case studies which informed the creation of the framework are:

- **Terrorgram:** A network of 208 accelerationist white-supremacist channels on Telegram, from which ISD gathered over 1,000,000 messages sent between 2016–2020;
- **Jack Reed:** Content identified in the prosecution of Jack Reed, a neo-Nazi and the youngest person to be prosecuted for terror offences in the U.K.;
- **Christchurch attack:** Content identified in the Royal Commission of Inquiry into the terrorist attack on Christchurch masjidain on 15 March 2019;
- **Caliphate Cache:** Content gathered from a pro-ISIS file storage site containing 2 terabytes of violent and non-violent Islamist content used by ISIS supporters to archive the ideology of the Islamic State; and
- **Islamist terror content:** Commonly shared content identified in the trials of U.K. Islamists convicted of terror offences.

Qualitative analysis of these diverse violent extremist environments allowed for the identification of cross-cutting categories of content as well as their distinct purpose(s) in the motivation, radicalization, and facilitation of violent extremist or terrorist activity, which formed the basis of the framework outlined below. As we will go on to explain, these distinct roles can be highly contextual, and any single item of content might appear innocuous in isolation.

Across the case studies we identified three broad categories for the classification of content, which will be unpacked in greater detail in the following section:

1. **Instructional material** which contains guidance on operational aspects of terrorist and violent extremist activity. This includes guidance on the manufacture and execution of attacks, as well as guides on combat drills, fitness, and non-violent activism such as sticker campaigning.

2. **Ideological material** that is designed to specifically further a violent extremist or terrorist world view. This includes key texts and lectures which provide the theoretical underpinning for a terrorist or violent extremist cause, and which provide an explanation for why the world is a certain way.

3. **Inspirational material** designed to reinforce a violent extremist or terrorist mind-set. This includes a wide range of content that is designed to elicit a reaction or response in the radicalized mind. This includes material intended to provoke hatred towards a particular group of people or promote pride and support for a particular cause. Notably, this category of content is the least well-defined in the existing literature.

Our analysis also revealed two further significant axes of division within the content in these violent extremist environments. Within the categories above, content is divided into explicitly violent and non-violent material, as well as classified between "official" group-based violent extremist material and content not associated with a formal organization. Therefore, in addition to the three content categories outlined above, our framework also incorporates the following distinctions for classifying content:

1. **Violent/non-violent content:** Violent extremist content includes material that depicts acts of violence, alludes to the preparation of violent acts, or is specifically designed to legitimize or inspire violent activity, as well as relevant extremist material that instead relates to another non-violent theme. This category reflects the ambivalent role of violent content within online extremist environments and offline radicalization pathways, recognizing that some platforms may want to focus counter-efforts more narrowly on explicitly violent content as a minimalist approach to tackling violent extremism online.

2. **Group/non-group content:** Reflecting the overall focus of this paper, this category relates to whether or not content has been produced by a particular violent extremist group. Although this framework is seeking to establish a group-agnostic taxonomy of terrorist or violent extremist content, we selected this for inclusion as branded material produced by violent extremist groups still played a major role in our analysis of ostensibly "post-organizational" spaces. When seeking to apply this framework, tech companies might consider "groups" in relation to specific proscribed organizations or adopt a broader approach, recognizing that certain tech companies already ban content from movements such as QAnon which are not proscribed.

Based on the different content categories outlined above, we constructed the following taxonomical framework, with content falling into one or several boxes:

	Inspirational material Content which can reinforce a violent extremist or terrorist mind-set		**Ideological material** Material which is specifically trying to further a violent extremist or terrorist world-view		**Instructional material** Content which contains instructions on operational aspects of terrorist activity	
Violent	Non-Group	Group	Non-Group	Group	Non-Group	Group
Non-Violent	Non-Group	Group	Non-Group	Group	Non-Group	Group

Figure 1. Classification framework for post-organizational violent extremist content

To help ground each of these subcategories more practically, sub-definitions were established to describe the different content types encompassed within the framework:

	Inspirational material Content which can reinforce a violent extremist or terrorist mind-set		**Ideological material** Material which is specifically trying to further a violent extremist or terrorist world-view		**Instructional material** Content which contains instructions on operational aspects of terrorist activity	
Violent	**Non-Group** Violent material not associated with a specific violent extremist group, designed to inspire violent extremism or terrorism, including copycat attacks.	**Group** Violent material group material designed to inspire violent extremism or terrorism.	**Non-Group** Material not associated with a specific violent extremist group, which makes the ideological case for extremist violence.	**Group** Violent extremist group material which makes the ideological case for violence.	**Non-Group** General instructional material providing operational guidance on carrying out acts of violence and terrorism.	**Group** Violent extremist group material providing operational guidance on conducting acts of violence and or terrorism.
Non-Violent	**Non-Group** Material not associated with a specific violent extremist group, which nonetheless inspires the mood music for violent extremism.	**Group** Violent extremist group material designed to inspire supporters, but not necessarily to violence.	**Non-Group** Material not associated with a specific violent group, which nonetheless helps to build the "system of meaning" underpinning a violent extremist ideology.	**Group** Violent extremist group material designed to build the "system of meaning" underpinning a violent extremist ideology.	**Non-Group** General instructional material providing operational guidance on non-violent activities relevant to violent extremism.	**Group** Violent extremist group instructional material providing operational guidance on non-violent activities.

Figure 2. Sub-definitions of post-organizational violent extremist content

	Inspirational	Ideological	Instructional
Violent	Group	Group	Group
	Non-Group	Non-Group	Non-Group
Non-Violent	Group	Group	Group
	Non-Group	Non-Group	Non-Group

Parameters of the framework, edge cases, and definitional challenges

When testing and developing this framework, we identified a number of definitional questions and instructive edge cases sitting at the borders of the categories, which are informative when considering its practical application. The permeability of the categories outlined in our framework and breadth of content contained within show the importance of a more sophisticated understanding of the interplay between different types of "organizational" and non-group-based content, as well as violent and non-violent material, contained within the wide constellation of online violent extremist communities.

Multi-category Content

It is recognized that certain documents, lectures, or sermons, particularly those which are of significant length, may sit across several of the above categories. For example, the 2011 Norway attacker's manifesto document contains instructional sections relating to the selection of targets for attacks, a wide range of ideological material designed to explain his world view, but also numerous statistics which help inspire hatred towards minority communities. To help circumnavigate this challenge we sought to establish through qualitative analysis the primary function for a piece of content.

Inspirational Versus Ideological Content

In distinguishing between inspirational and ideological violent extremist material, this framework first separates out content intended to reinforce a violent extremist or terrorist mind-set, and then material that tries to further a violent extremist or terrorist worldview.

This is an important distinction, recognizing that some violent extremist propaganda seeks to imbue its audiences with what Haroro Ingram describes as a "competitive system of meaning" which acts as a lens through which supporters are compelled to perceive and judge the world,[138] as opposed to texts which do not engage in ideology construction per se but rather seek to motivate and inspire (for example, by bolstering in-group identity or emphasizing the urgency of crises which "require" extremist solutions). In reality, a considerable proportion of the content analyzed in the case studies below serves both an inspirational and ideological purpose, designed to both consolidate an extremist worldview and encourage adherents to mobilize around it. However, while it is perfectly viable (and indeed to be expected) that content can serve multiple violent extremist ends, it is likely useful when classifying violent extremist content – and considering proportionate responses – to assess whether it is primarily serving an ideological or an inspirational purpose, when it is clearly not instructional in nature.

The Thresholds of Inspirational Content

When defining **inspirational** material which reinforces a violent extremist mind-set, the importance of the intent behind the creation of a piece of content became apparent – whether material was designed to reinforce a terrorist or violent extremist mind-set, or whether it incidentally reinforces extremist narratives. This distinction is evident when you explore the range of seemingly innocuous material which is referenced, shared, and promoted by violent extremists and terrorists. For example, the manifesto document produced by the Oslo attacker contains references and quotes from articles written by a wide range of figures including British journalists Melanie Phillips and Jeremy Clarkson.

In these instances, the source material referenced was clearly not originally designed to inspire violent extremism, and therefore should not be seen as inherently inspirational. However, it was nevertheless repurposed in a way that was intended to reinforce a violent extremist mind-set. This corpus of material that has the potential to inspire violent extremist activity is vast – in the case of the extreme right wing this could include for example official government figures relating to the changing ethnic make-up of a country and material which is critical of immigration. Meanwhile, religious extremists might selectively quote scriptural sources out of context to support their ideology. Accordingly, when testing and applying this framework in the **case studies** outlined below, we primarily focused categorization on material where an intent to inspire violent extremism can be demonstrated or adequately inferred. We recognize this can present a challenge for classification, but believe it is nevertheless crucial for increasing the precision of this framework and decreasing "creep" toward categorizing otherwise innocuous content which has been instrumentalized for violent extremist purposes.

[138] Haroro Ingram, "Deciphering the Siren Call of Militant Islamist Propaganda: Meaning, Credibility & Behavioural Change," ICCT, The Hague, September 2016, link

Ambiguity in Distinguishing Between Group and Non-group Content

It may not always be easy to make clear distinctions between group and non-group content. In some cases, specific pieces of content may have been created by individuals who are closely associated with a specific movement. While David Duke and his work are often celebrated in extreme right circles, his name is intimately tied to his former position as the grand wizard of Ku Klux Klan. Similarly, while the Salafi-jihadist cleric Anwar al-Awlaki ultimately became associated with al-Qaeda in the Arabian Peninsula (AQAP), much of the material he produced preceded that association. Similarly, violent extremist groups may reference content by individuals as a source of inspiration without these individuals having a firm affiliation with the group. In determining whether content was "group" or "non-group" produced, we thus paid close attention to whether content was specifically affiliated with a group – for example through branding or direct support for a violent extremist organization. In the case of content that was n0t branded, we also chose to include texts by individuals with leadership roles in movements as "group" produced if this content was made at the time the individual was affiliated with the group.

Understanding the Role of Non-violent Content

The creation of this framework is specifically intended to nurture a better understanding of violent extremist activity; however, we nevertheless include non-violent content within our framework. This is justified through reflection on the fact that violent extremism does not just relate to violent actions, but also to broader activity that seeks to dehumanize and delegitimize societal out-groups, as well as a wide range of non-violent efforts to build and strengthen movements and radicalize individuals. Recent post-organizational terrorism highlights the blurred and ambiguous boundaries between so-called "violent" and "non-violent" extremism and necessitates the inclusion of non-violent content in this taxonomy. For example, the Identitarian movement and the "great replacement" theory were particularly influential in inspiring and providing the ideological underpinning of the 2019 Christchurch terrorist attack despite being nominally non-violent.

Applying the framework to "real-world" cases

The following case studies apply this prototype taxonomy to real-world instances of terrorism. They shine a light on the diverse range of online content which underpins terrorist and violent extremist activity.

Case study: "Terrorgram"

The encrypted messaging platform Telegram represents a major online hub for contemporary violent extremist activity.[139] White-supremacist groups have proliferated on the platform, forming a network of public channels that has been referred to as

[139] Michael Schwirtz, "Telegram, Pro-Democracy Tool, Struggles Over New Fans From Far Right," New York Times, January 26, 2021, link

"Terrorgram."[140] Previous ISD research found over two hundred pro-terrorist channels glorifying terrorism, calling for violence, spreading extremist ideological material, and demonizing minority groups without having any formal affiliation with a specific group, lending itself well to a case study in post-organizational violent extremist mobilization.[141]

	Inspirational material Content which can reinforce a violent extremist or terrorist mind-set		**Ideological material** Material which is specifically trying to further a violent extremist or terrorist world-view		**Instructional material** Content which contains instructions on operational aspects of terrorist activity	
	Non-Group	**Group**	**Non-Group**	**Group**	**Non-Group**	**Group**
Violent	Oslo Manifesto CHCH Manifesto Charleston Manifesto Saint Memes	AWD Videos Azov Videos The Base Serbian millitary videos and songs Calls for violence Glorification of historical fascists	Turner Diaries Siege	Hitler Codreau O9A books	Adivse for firearm construction	National Action manuals
Non-Violent	Antisemitic memes White Genocide memes Aethetic images of European beauty Fashwave White supremacist rock music Images of degeneracy Christian / pagan symbols	N/A	Books by Julius evola and David Irving	Jewish Supremacism by David Duke	Images of target buildings	N/A

Figure 3. Examples of content identified in "Terrorgram"

Through our analysis of the network, we found a wealth of content fitting each of our overarching content categories. In particular, we found considerable **inspirational** material shared across the channels analyzed. This included a large amount of violent content produced by **groups** such as Atomwaffen and The Base, as well as manifesto documents produced by **lone actor** terrorists and material glamorizing individuals who have committed attacks.

However, the largest amount of content identified notably fit the category of **non-group non-violent inspirational material**. This included white-supremacist music and a vast array of user-created memes which convey racist, antisemitic, and misogynist ideas

[140] Hope Not Hate, "The Terrorgram Network: A spiral towards bloodshed," Hope Not Hate, published 2021, link

[141] Jakob Guhl and Jacob Davey, "A safe space to hate: White supremacist mobilisation on Telegram," ISD, June 2021, link

or celebrating extreme right ideologues. Beyond material which more obviously was designed by and for extremists, we also identified a vast amount of cultural material, and material relating to sex, gender, and the family. This cultural material included historical photographs, photographs of "traditional" looking beautiful women, and pictures of classical art – at times superimposed with inspirational slogans designed to reinforce a white-supremacist world view, such as "embrace tradition, reject modernity" (which themselves are common features of fascist rhetoric as identified by Umberto Eco).[142]

Importantly, such apparently innocuous images were shared in the context of communities that actively advocate for extreme violence, illustrating the role which non-violent content can play in reinforcing a violent extremist mind-set, and suggesting that such material could be an indicator in certain circumstances of more concerning activity within a community.

Groups also commonly shared long-form **ideological** content including PDFs and audio recordings of books by a range of ideologues and key antisemitic texts such as the "Protocols of the Elders of Zion" forgery. Additionally, we encountered a range of **instructional** material including guerrilla warfare manuals that include training and equipment recommendations, videos providing advice on firearm construction or images of religious buildings that could serve as potential targets, as well as more innocent-looking material such as advice on self-sufficiency.

Although some of the channels identified were devoted to one specific type of content (such as several devoted to pictures of attractive and apparently fascist women), many contained mixtures of the content types outlined above. This suggests that post-organizational violent extremist communities draw from a wide range of content types, and raises questions around whether a narrow focus on particularly violent material is sufficient in capturing the reality of contemporary violent extremist activity online.

Case Study: Jack Reed

In December 2020 the teenager Jack Reed became the youngest individual convicted for terrorism-related offences in the United Kingdom.[143] Reed's online searches and materials he possessed were used as evidence in his trial. This case study outlines some of the materials that may have inspired Reed, helped shape his ideology, and informed his preparations to attack targets in the Durham area as well as inspirational and instructional content Reed produced himself.

Reed's case closely follows a post-organizational paradigm. He self-radicalized largely through online engagement with extremist communities, was not a member of a proscribed terrorist organization, and at the time of his arrest was preparing for a lone actor act of terrorism. Accordingly, this case study is particularly useful for the purposes of testing this

[142] Umberto Eco, "Ur-Fascism," New York Review of Books, June 22, 1995, link
[143] John Simpson, "Neo-Nazi teenager Jack Reed plotted to attack his mother's office," The Times, January 16, 2021, link

framework.

	Inspirational material Content which can reinforce a violent extremist or terrorist mind-set		**Ideological material** Material which is specifically trying to further a violent extremist or terrorist world-view		**Instructional material** Content which contains instructions on operational aspects of terrorist activity	
Violent	**Non-Group** Charleston Manifesto Reed's manifesto / manual	**Group** Field Manual (FM) 6-2003 Ethnic Cleansing Operations	**Non-Group** Fascist Forge usage Siege	**Group** Historical fascist literature (Hitler, Speer and Codreanu) Texts by AWD, O9A and Tempel ov Blood	**Non-Group** Reed's to-do list Online-searches for material on firearms, explosives, ammunition and weapons Including 21 Techniques of Silent Killing by Hei Long Homemade C4 by Ragnor Benson Big Book of Mischief by David Richards	**Group** Atomwaffen Manual
Non-Violent	**Non-Group** Antisemitic, homophobic and Neo-Nazi memes Selfie w. Tommy Robinson	**Group** N/A	**Non-Group** The Way of Men by Jack Donovan The Synagogue of Satan by Andrew Carrington Hitchcock The Myth of German Villainy by Benton L. Bradbury	**Group** White Power by George Lincoln Rockwell	**Non-Group** Online-searches for Synagogue details	**Group** N/A

Figure 4. Examples of content identified in the Jack Reed case

An analysis of Reed's online activity and material in his possession that was revealed in his trial demonstrates how he was likely influenced by a wide range of content types. This included a diverse range of inspirational, ideological, and instructional material, including hateful memes, a range of ideological core texts, and instructions around making and detonating explosives.

Interestingly, some of this material crosses over with some of the content identified in the "Terrorgram" case study, suggesting a set of core texts which are particularly influential to post-organizational white-supremacist terrorism. Accordingly, producing lists of books that repeatedly reappear in convictions could be helpful when considering ways to bolster platform moderation efforts. Additionally, this would suggest that expanding current hashing technology beyond images and videos could help identify particularly important violent extremist material.

Case Study: Christchurch attack

In December 2020, the Royal Commission of Inquiry into the March 2019 terrorist attack in Christchurch was published.[144] The 800-report provided an in-depth assessment of the attackers' background and radicalization pathways. Crucially for this briefing, the report refers to a range of online materials and activities the terrorist engaged in leading up to the attack.

	Inspirational material Content which can reinforce a violent extremist or terrorist mind-set		**Ideological material** Material which is specifically trying to further a violent extremist or terrorist world-view		**Instructional material** Content which contains instructions on operational aspects of terrorist activity	
Violent	**Non-Group** Oslo Manifesto Christchurch Manifesto Christchurch Facebook page image collection	**Group** Azov Battalion Mein Kampf	**Non-Group** Turner Diaries Comments	**Group** True Blue Crew Lads Season Two group comments	**Non-Group** Oslo manifesto	**Group** N/A
Non-Violent	**Non-Group** Black Sun Symbol	**Group** 14 words (The Order)	**Non-Group** "Great Replacement" by Renaud Camus Black Sun Symbol The Decline of the West by Oswald Spengler A Short History of Decay by E. M. Cioran	**Group** N/A	**Non-Group** Footage of Masji an-Nur	**Group** N/A

Figure 5. Examples of content identified in the Christchurch inquiry

Central to the Christchurch attacker's radicalization trajectory was **violent non-group content** such as the Oslo manifesto, which played a crucial inspirational and ideological role. This also served a more instructional purpose, with the Commission report showing that the actions of the Oslo attacker provided a blueprint for the Christchurch attacker during the planning stages. This included activities such as joining a gym, using steroids, practicing his rifle skills, manifesto writing as well as publication timing and operational security (e.g. digital hygiene). While some of the sub-sections of the Oslo manifesto may appear innocuous in isolation, they were designed as a coherent call for violent extremism when looking at the entire body of the text. It is therefore worth platforms considering whether quotations from violent extremist manifestos should be treated as terrorist content, even if the content in question appears innocent when separated from its original text.

[144] Royal Commission of Inquiry into the terrorist attack on Christchurch masjidain on 15 March 2019, "Ko tō tātou kāinga tēnei," Royal Commission of Inquiry into the terrorist attack on Christchurch masjidain on 15 March 2019," December 2020, link

The Christchurch attacker possessed a range of **ideological literature** that was either explicitly violently extremist in nature or is widely read by the extreme right, even though it was not produced for this purpose. In addition to the “Great Replacement” by far-right French philosopher Renaud Camus, which was referenced in the title of the attackers’ manifesto, the Commission report lists works by Oswald Mosley, Oswald Spengler, and E. M. Cioran as books purchased by the Christchurch attacker. In comments posted on the United Patriots Front Facebook page, the Christchurch attacker expressed hope he would be taking part in their execution on the “Day of the Rope,” a reference to The Turner Diaries, a novel about a race war that has inspired numerous major violent attacks in the decades since its publication in 1978. On the same Facebook page, the attacker makes clear he had engaged with the writings and ideas of Adolf Hitler. Referring to **Mein Kampf**, the attacker argued that perceived victimhood and political grievances of whites should be the center of any political communication aimed at recruiting new followers. While there can be a legitimate interest in historical fascist literature, this could perhaps be distinguished from discussions that take positive inspiration from such works or advocate for the ideas expressed in it.

The white-supremacist Azov Battalion in Ukraine also appears to have been a source of **violent group-based inspiration** for the Christchurch attacker. According to the Royal Commission, the mother of the Christchurch attacker became extremely worried her son was about to move to Ukraine in order to fight for Azov. The attacker also appeared to have been attracted to Ukraine due to its cheaper cost of living and more ethnically homogeneous society. It therefore appears likely he would have watched Azov propaganda material, which is widely circulated among extreme right communities online.

Important non-violent instructional content during the preparatory stages for the Christchurch attack was a video from inside Masjid an-Nur, which the attacker had taken from a public Facebook page and saved to his phone four days before the attack. The individual who posted the video was not known to the attacker or affiliated with the far-right. This suggests that even entirely innocuous content created for benign purposes may contain information that could be relevant for the preparatory stages of a terrorist or violent extremist attack. Identifying what types of innocuous content may contain such relevant information and should raise flags in combination with more overtly violent extremist content presents a major challenge.

Case Study: Caliphate Cache

In the wake of Abu Bakr al-Baghdadi's death in October 2019, ISD researchers discovered a network of ISIS supporters sharing links to one of the largest known online repositories of ISIS material. Provisionally analyzed by researchers from ISD and West Point's Combating Terrorism Center,[145] this two terabyte self-contained "archive" represents an important case study of the breadth of violent extremism related material collected by ISIS supporters who are not formally affiliated with the group. Its post-organizational dimension is derived from the broad range of both group-based as well as non-official "supporter" material and predecessor content, spanning ideological, inspirational, and instructional typologies, as well as many items which straddle these categories.

	Inspirational material Content which can reinforce a violent extremist or terrorist mind-set		**Ideological material** Material which is specifically trying to further a violent extremist or terrorist world-view		**Instructional material** Content which contains instructions on operational aspects of terrorist activity	
Violent	**Non-Group** Global Islamic Resistance Call - Abu Musab al-Suri (while al-Suri was linked, the work itself has been promoted by a range of groups, including those inclined to violence and those that are not)	**Group** 'Archives of the Martyrs'	**Non-Group** 'Scholars of Jihad' (al-Maqdisi) Unofficial pro-ISIS support groups	**Group** 'Scholars of Jihad' (Zarrouk, al-Awlaki, al-Anbari) 'Emirs of Jihad' ISIS 'fatwas over the airways'	**Non-Group** Management of Savagery - Abu Bakr al-Naji 'Mujahid's Bag'	**Group** 'Archive of Military Science' 'Four Easy Ways...' Al Saqri Media Collection
Non-Violent	**Non-Group** Nasheeds Poetry	**Group** Poetry Essays of Eid and life in "Wilayat" of the "Caliphate"	**Non-Group** Tafsir by al-Maqdisi Suliman al-Alwan Abdulaziz al-Tarefe	**Group** N/A	**Non-Group** Tactical guides for evading security	**Group** Hamas manual on fooling prison authorities, e.g. evading jailhouse snitches

Figure 6. Examples of content identified in the Caliphate Cache

[145] Moustafa Ayad, Amarnath Amarasingam, and Audrey Alexander, "The Cloud Caliphate: Archiving the Islamic State in Real-Time," Combating Terrorism Center at West Point and ISD, May 2021, link

When looking at the spread of content across the framework, it is perhaps unsurprising that the vast majority of material sits in the violent classification. This site appears to be designed as a centralized digital repository for Salafi-jihadi true believers rather than a means of engaging those on the ideological fence. However, the archive notably contains a deep bank of digital material which can be deployed in order to inspire, instruct, and even educate would-be supporters. Analyzing the relative size of "folders" within the Cache can also provide rough indications about the scale of content intended for different inspirational, ideological, or instructional purposes.

One of the larger folders, "Scholars of Jihad," contains material that skews towards the **ideological**, hosting content from a mixture of Salafi-jihadi ideologues formally affiliated with ISIS or its predecessor groups as well as a number of those who share an ideology but not a group affiliation. These include group unaffiliated "scholars" such as Abu Muhammad al-Maqdisi, the al-Qaeda affiliated Anwar al-Awlaki, and ISIS ideologues. Meanwhile, the parallel "Emirs of Jihad" folder includes figures such as Abu Hamza al-Quraishi who are more explicitly group affiliated, but also serve an inspirational as well as an ideological purpose.

The Cache contains a relatively comprehensive archive of "official" ISIS material. However, this stands in considerable contrast to content such as the "Management of Savagery," a seminal text written in 2004 by Abu Bakr Naji, which provides much of the instructional and ideological roadmap for global Salafi-jihadism but is not associated with any specific terrorist group. Meanwhile, Tafsir – commentary and scholarly interpretation of every verse in the Quran – hosted in the archive from scholars amenable to the Salafi-jihadi cause provides important non-group-based ideological underpinnings to ISIS. There are 53 non-official ISIS-linked groups archived in the Cache within a folder dubbed "Support Organizations." These groups are all individually branded, and the material could **either be classified as non-group or group** content depending on how it is defined. The status of these groups represent a critical missing piece of the puzzle in identifying and classifying branded terrorist support materials across platforms.

Instructional content in the Cache includes formal violent operational texts such as the "Archive of Military Science" and "Four Easy Ways" series by the ISIS media agency al-Saqri, which includes guidance on making explosive belts. The Cache also includes user-generated folders of assorted operational documents such as "The Mujahid's Bag" as well as a large folder on operational security which includes a range of group and "supporter" content not clearly linked to ISIS.

Finally, **inspirational materials** include a comprehensive "Archives of the Martyrs" which stretches back to the death of Abu Musab al-Zarqawi in 2006. More ambiguous from a classification perspective are numerous folders containing nasheeds (i.e. religious anthems), which seek to provide the literal "mood music" for jihadi violence but are not necessarily explicitly associated with ISIS as a group. In a similar vein, a large collection of jihadi poetry in the archive contributes to the cultural backbone for the Salafi-jihadi "system of meaning" while not being explicitly associated with the group or even explicitly inciting violence.

Case Study: U.K. Islamist terrorism cases

The following case study draws on an amalgamation of content from across several U.K. terrorism cases published in an article by Donald Holbrook.[146] The repetition of key ideologues and materials is a useful way to understand not only what specific content speaks to the Islamist terrorist mind-set, but how old and new resources intersect to justify violence in a modern context.

	Inspirational material Content which can reinforce a violent extremist or terrorist mind-set		**Ideological material** Material which is specifically trying to further a violent extremist or terrorist world-view		**Instructional material** Content which contains instructions on operational aspects of terrorist activity	
Violent	**Non-Group** Babar Ahmad / Azzam publications: In the Hearts of Green Birds (aka: Martyrs of Bosnia)	**Group** ISIS videos Al-Qai'da videos: speeches by Osama bin Laden	**Non-Group** Awlaki: The Book of Johad (also the original Ibn Nuhaas version) Awlaki: Allah is preparing us for vicotry Azzam: Join the Caravan Azzam: Defence of the Muslim Lands: The First Obligation after Iman Azzam: The signs of Allah the Most Merciful Ar-Rahmaan in the Jihad of Afghanistan Awlaki: Allah is Peparing us for victory Awlaki: Constants on the Path of Jihad Azaam: The Lofty Mountain	**Group** Inspire magazine (al-Qai'da) Dabiq Magazine (ISIS)	**Non-Group** N/A	**Group** Inspire magazine (al-Qai'da) Dabiq Magazine (ISIS)
Non-Violent	**Non-Group** Awlaki: The Life of Mohammed (both Makkan and Median Periods) Awlaki: The Hereafter / The Afterlife / Al-Achira Awlaki: Lives of the Prophets Abu Muhammad Al-Maqdisi: This is our Aqidah	**Group** ISIS videos: state building; camaraderie	**Non-Group** Awlaki: Umar Ibn al Khattab - His life and times Babar Ahmad / Azzam Publications: Under the shades of the swords Sqyyid Qutb: In the shade of the Quran Sayyid Qutb: Milestones	**Group** ISIS videos: legitimacy of the group, illegitimacy of other Islamist groups	**Non-Group** N/A	**Group** N/A

Figure 7. Examples of content found in U.K. Islamist cases

[146] Donald Holbrook, "The Spread of its Message: Studying the Prominence of al-Qaida Materials in UK Terrorism Investigations," Perspectives on Terrorism 11, no. 6, (2017): 89–100, link

One of the challenges of clearly delineating Islamist terrorist content is the volume of inspirational material that has been produced. Much of it is not explicit in any calls for violence, but for those determined on joining the violent jihadist cause, there are veiled allusions to the legitimacy of violence, often rooted in Islamic history.

Even with the more recent wealth of ISIS propaganda which included more obvious calls to violence, the overwhelming majority of Islamist terrorist content found in U.K. terrorism investigations are the enduring historical and theological justifications for jihad and given a refreshed "lease on life" by al-Qaeda. Consequently, when analyzing the presence of Islamist propaganda, the authors (and sometimes narrators) of the materials are as important to consider as the titles and content. Similarly, publishers become an important feature of this content, with as-Sahab Media, Azzam Publications, and the Maktabah al-Ansar bookshop being most notable, along with ISIS's more recent Amaq News Agency.

The nature of Islamist texts leads them to be relevant to both inspirational and ideological content, rooted as they are in Islamic history and theology. Anwar al-Awlaki is popular for his lectures that bring Islam's history and that of the prophet Muhammad and his companions to life. They are inspirational, not in relation to violence, but in how they cement Awlaki's credibility as a scholar of Islam, which in turn makes his radicalizing content more compelling to his audience. While the inspirational texts do contain some endorsements of violence, they could not be construed as a direct incitement to jihad.

In terms of violent extremist ideology, the majority comes from just a small handful of ideologues. Once again, Awlaki is the most prominent, and in his later years his content became more militant and explicitly endorsed violence. However, a range of other individuals also featured throughout the cases Holbrook analyzed, including Abdullah Azzam, as well as the joint ideological/instructional content of al-Qaeda's Inspire and ISIS's Dabiq magazines. These magazines have been popular due to their framing of the ideological justification for terrorism, fetishizing and canonizing killed terrorist "martyrs," and offering step-by-step practical guides for building bombs, selecting targets, and conducting attacks.

It is clear from the materials in this section that authorities, authors, and ideologues continue to inspire even after they are dead. Material by Awlaki is still popular even though he was killed in September 2011. Abdullah Azzam, killed decades earlier, also continues to resonate, albeit via more contemporary gatekeepers, and the Saudi-born commander of Chechen militants "Ibn al-Khattab" has achieved iconic status. Islamist terrorists do not just consume recently created content, but draw ideological succor from a long and varied literary tradition.

As far as established terrorist organizations are concerned, al-Qaeda seems to be more prominent than ISIS (or any of its offshoots around the world), but materials from al-Qaeda do not feature in isolation. People seeking to become involved in terrorism collect a broad repertoire of media publications conveying religious, political, and ideological content from a variety of different sources and publishers. No single group, cluster, or school appears to dominate, and the range of influencers is diverse. Attempts to flag and respond to this content should be able to cross-reference the multiple ideologies and sources to determine a growing interest in understanding the permissibility or even obligation of violence.

Using this framework for cross-platform efforts to respond to violent extremist content

Part of the rationale for the development of the framework laid out in this paper is to find an approach to defining and labeling violent extremist material that might be useful to a variety of technology platforms. Platforms define groups and content relating to violent extremism differently, as well as diverging in the actions taken in response to these groups or materials on their services. The utility of a framework therefore partly lies in whether it can adequately capture some crossover between platform policies that will inevitably continue to differ in order to streamline responses to cross-platform violent extremist activity. This could therefore apply to platforms' application of the hash-sharing database, but could also provide considerations relevant to crisis-response decision making or individual instances of platform collaboration tackling new iterations of violent extremist threats across services.

A framework can also be useful to help identify gaps in individual platform policies in the context of an ever-changing violent extremist online environment. By assessing if and how current platform policies are covered by the conceptual framework laid out here, we can start to ascertain where types of violent extremist activity are currently described by platform policies.

To start to understand if and how such a framework might be useful despite the complications of individual platform policies and responses, it is possible to plot existing individual company policies against the framework to start to see commonalities and gaps.

As an illustrative exercise, we use the current policies of Facebook, Twitter, and YouTube as a sample to identify some initial ways the framework can provide insights into the types of violent extremist activity that are largely covered by most companies' existing policies, covered by some but not others, or currently largely outside of companies' terms of service or community guidelines.

	Inspirational material Content which can reinforce a violent extremist or terrorist mind-set		**Ideological material** Material which is specifically trying to further a violent extremist or terrorist world-view		**Instructional material** Content which contains instructions on operational aspects of terrorist activity	
Violent	**Non-Group** Twitter's "Violent threats policy" Twitter's "Glorification of violence policy" YouTube's "Harmful or dangerous content" policy YouTube's "Violent or graphic content" policy	**Group** Facebook's "Dangerous individuals and organisations" policy Twitter's "Violent extremist groups and violent organisations" policy YouTube's "Violent criminal organisations" policy	**Non-Group** Twitter's " Violent threats policy" Twitter's "Glorification of violence policy" YouTube's "Harmful or dangerous content" policy YouTube's "Violent or graphic content" policy	**Group** Facebook's "Dangerous individuals and organisations" policy Twitter's "Violent extremist groups and violent organisations" policy YouTube's "Violent criminal organisations" policy	**Non-Group** Facebook's "Coordinating harm and publicising crime" policy Twitter's "Violent threats policy" Twitter's "Glorification of violence policy" YouTube's "Harmful or dangerous content" policy YouTube's "Violent or graphic content" policy	**Group** Facebook's "Dangerous individuals and organisations" policy Twitter's "Violent extremist groups and violent organisations" policy YouTube's "Violent criminal organisations" policy
Non-Violent	**Non-Group** N/A	**Group** Facebook's "Dangerous individuals and organisations" policy Twitter's "Violent extremist groups and violent organisations" policy YouTube's "Violent criminal organisations" policy	**Non-Group** N/A	**Group** Facebook's "Dangerous individuals and organisations" policy Twitter's "Violent extremist groups and violent organisations" policy YouTube's "Violent criminal organisations" policy	**Non-Group** Facebook's "Coordinating harm and publicising crime" policy	**Group** Facebook's "Dangerous individuals and organisations" policy Twitter's "Violent extremist groups and violent organisations" policy YouTube's "Violent criminal organisations" policy

Figure 8. Areas of the framework currently covered by platform policies

Here it remains likely that the policies outlined above will only partially cover content in a particular sub-category of the framework. For example, with **instructional content** it will not always be immediately obvious that it is tied to violent extremism and therefore may not be covered by Facebook's "coordinating harm and publicizing crime" policy.

With this caveat in mind, we can see that most explicitly **violent activity** is fully or partially covered by the three platforms used in this test, as is most **group activity**. Yet as the analysis detailed above illustrates, not all violent extremism is driven by "dangerous organizations" but also looser digital communities and content groups. The areas that platform policies do not appear to currently cover are **non-violent non-group** material. However, this content appeared across all of the case studies analyzed above.

This content is the most challenging to develop moderation policies around. Much of it relates to the use of symbols, memes, literature, and cultural material where links to violent extremism are not always explicit or apparent without expert understanding of violent extremist mobilization and ideology. Yet it is clearly an important cornerstone of violent extremist communications online, and a persistent presence in contemporary post-organizational violent extremism and terrorism. Creating responses to such material will thus require responses that go beyond individual pieces of content and instead attempt to interrogate the intention behind their circulation and the behaviors of communities involved in this.

Current definitions of terrorist content do not cover such content and instead focus on material that is either associated with proscribed organizations or explicitly with the incitement or preparation of terrorist offences and violence. These include those proposed by the European Commission,[147] the European Parliamentary Research Service,[148] the U.K. Government,[149] and Digital Europe[150] (the trade association representing digital industries in Europe). Nor is such content covered by GIFCT's hash-sharing database.[151] However, such material is covered in broader analyses of **extremist** content online, such as **Holbrook's Extremism Media Index** where it is presented as "fringe material" which exists on a spectrum with violent content.[152] It is (likely) similarly covered in part by individual platform policies governing hate speech and harassment.

This poses the question if it is helpful or desirable to incorporate non-violent non-group created material into broader conceptions of terrorist activity. Securitizing non-violent ideological content such as Camus's **The Great Replacement** or non-violent inspirational material such as memes and cultural references is doubtlessly problematic. To approach such content in a ham-handed fashion lacking nuance would almost certainly bring with it the prospect of banning legitimate content and have a disastrous impact on freedom of speech. However, completely isolating such content from categorization frameworks addressing terrorist activity denies the key role it plays in radicalization and movement-building by violent extremists and terrorists, and negates the fact that such material is used as circumstantial evidence in the prosecution of terror offenders. Indeed, in the current post-organizational landscape where self-radicalization and fluid online coordination are superseding more structured group-based dynamics, the role of key texts and material which conveys core violent extremist tropes is arguably more important than ever. This itself raises further questions around whether the removal of content is the only

[147] Joris van Hoboken, Vrije Universiteit Brussels and University of Amsterdam, "The Proposed EU Terrorism Content Regulation: Analysis and Recommendations with Respect to Freedom of Expression Implications," Transatlantic Working Group, May 3, 2019, link
[148] European Parliamentary Research Service, "Addressing the dissemination of terrorist content online," April 9, 2021, link
[149] Home Office, "Interim code of practice on terrorist content and activity online," December 2020, link
[150] Microsoft Corporate Blogs, "Microsoft's approach to terrorist content online," May 20, 2016, link
[151] GIFCT, GIFCT Transparency Report.
[152] Donald Holbrook, "Designing and Applying an 'Extremist Media Indix'." Perspectives on Terrorism 9, no. 5, October 2015, link

appropriate tool when addressing activity that is relevant to violent extremism and opens up broader possible approaches.

Accordingly, should a company decide to expand their policies to include the broader types of content associated with violent extremism identified in this briefing, there is an absolute need for definitional clarity and the parameters between potentially dangerous material and free speech. Finding ways to effectively implement this framework in a sensitive but robust fashion will be key in negating concerns around infringement on civil liberties. Such technical operationalization will require careful planning beyond the scope of this briefing; however, to help inform any future implementation of this taxonomy, we have identified the following approaches which could shape content moderation efforts:

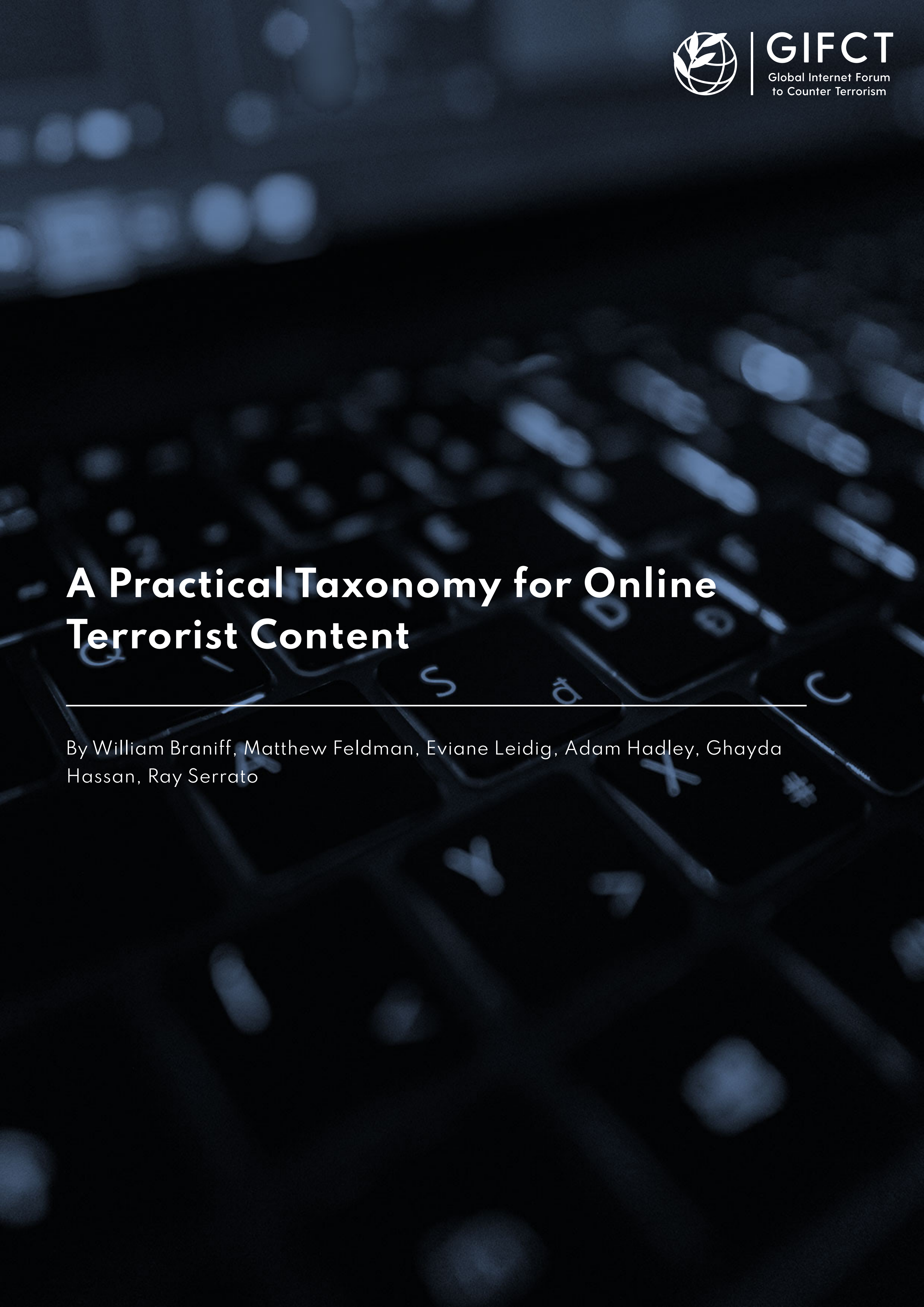
- **"Relevant" content flags:** A tiered approach could be adopted where non-violent non-group content which is circulated by terrorist and violent extremist communities and individuals is flagged as potentially relevant. Here, when an account or community is removed from a platform for terrorist or violent extremist activity, the broader content they shared would be flagged for future reference but would not be subject to automatic or immediate removal. This would mean that should such content be shared online in the future, the communities and users sharing this content could be marked for further investigation, with tiers of potential risk ascribed to the volume of potentially relevant content shared. In essence, this raises the possibility of creating a hash-sharing database that is not entirely focused on the takedown and removal of content and instead using technology to inform subtler approaches to countering online mobilization by violent extremist communities.
- **High-risk content flags:** Where non-violent non-group-based content is commonly featured in the conviction of terrorists, such material could be flagged as "potentially high-risk." The appearance of this content could thus be used to identify an online community or account as sharing material that may be relevant to violent extremism and terrorism and worthy of further investigation. Accordingly, an academic discussion group focused on French new-right philosophy sharing The Great Replacement would be treated differently from a community or individual promoting ethno-nationalism by sharing the same text.
- **Alternative approaches to content:** The hash-sharing database is currently designed to inform the removal of terrorist content. However, raised above is the possibility of creating hashes of potentially risky or insightful content that does not cross the threshold for immediate removal. Such hashes could be used to inform other less-blunt approaches to content that go beyond takedowns, including removing communities hosting large quantities of potentially high-risk content from recommendation systems.
- **Behavior-sensitive moderation efforts:** To supplement purely content-based approaches, platforms could seek to analyze the broader activity of communities regularly sharing non-violent non-group material. This could include the extent to which these communities are networked with other extremist hubs on the same or

other platforms or whether users appear to be deliberately radicalizing other users of a platform by (for example) spamming large quantities of potentially violent extremist content or directing users to other violent extremist communities.

- **The possibility of cross-platform moderation:** In the case studies of the Christchurch attack and Jack Reed discussed above, both individuals were operating across multiple social media platforms and forums. There are clearly limits to how much a company can do with regard to activity outside of its platform. However, this does raise the suggestion that companies should consider approaches to communities using their platforms that are directing people to insightful or risky content elsewhere.

Conclusion

Based on our analysis of five real-world case studies, we have created a prototype framework for the classification of violent extremist content in a way that is group-agnostic. This framework recognizes the nuances of violent extremist content online and helps advance beyond narrow conceptions which emphasize the role of specific organizations, branded terrorist material, and explicitly violent content in the inspiration of violent extremism. Our framework broadens these conceptions and recognizes the influential role that non-violent content like user-created memes and key ideological texts can have. However, through the analysis of real-world case studies, we can see that such content may appear innocuous or falls under protected speech. Accordingly, implementing this framework will require a careful approach that is conscious of fundamental rights and involves consideration of policy approaches that are more nuanced than the blanket removal of content.



GIFCT
Global Internet Forum
to Counter Terrorism

A Practical Taxonomy for Online Terrorist Content

By William Braniff, Matthew Feldman, Eviane Leidig, Adam Hadley, Ghayda Hassan, Ray Serrato

A Practical Taxonomy for Online Terrorist Content

By William Braniff, Matthew Feldman, Eviane Leidig, Adam Hadley, Ghayda Hassan and Ray Serrato

Overview

This work product is in response to the Global Internet Forum to Counter Terrorism's (GIFCT) Request for Proposals for taxonomic frameworks for terrorist content online.[153] The joint authors of this text – consisting of international specialists and GIFCT stakeholders convened by the Independent Advisory Committee (IAC) – have provided a practical resource in the form of a grid for use by online platforms to identify different types of terrorist content (X-axis) and online platform functions (Y-axis). Following this introduction, the sections below provide the context and assumptions underpinning this work product, a description of how GIFCT members can use this tool, indicative characterizations used in the resource, and a concluding section providing both recommendations and resources.

The impetus for this work originates with the 2019 Christchurch Call in response to the horrific livestream murder of Muslim worshippers in New Zealand on 15 March 2019. Although GIFCT had been founded two years earlier as a consortium of leading online platforms, governments, academics, and specialists coming together "to prevent terrorists and violent extremists from exploiting digital platforms," the events at Christchurch made such an initiative even more urgent. To address this dire need, the authors of this document aim to provide an easy-to-use resource for combating terrorism and violent extremism that is specifically oriented toward the following priorities:

- Help GIFCT member companies to better understand the violent extremist landscape they are navigating;
- Empower GIFCT member companies to make better decisions about content moderation, resource allocation, and information sharing;
- Help improve GIFCT member companies' Terms of Service by considering the spectrum of harmful content produced by terrorists and violent extremists beyond explicitly violent content;
- Help improve existing hash-sharing databases and other content moderation tools beyond the use of images and videos;
- Present a practical work product that addresses terrorism and violent extremism, along with key recommendations to aid companies in decision making on terroristic content.

[153] Critical assistance in the preparation and drafting of this document was also provided by GIFCT Independent Advisory Committee members from Ghana and the United Kingdom. The authors also wish to thank Milo Comerford at the Institute for Strategic Dialogue for insights and advice on the compilation of this document, as well as GIFCT stakeholders Bjørn Ihler, Johannah Lowin, Nayanka Paquete Perdigao, Nick Rasmussen, Dr. Erin Saltman and Thomas Thorley.

However, several caveats must be initially noted. First, given the challenges in defining terrorism, this work product does not seek to provide a static definition of terrorism or terrorist content. Instead, focus is placed upon key characteristics of violent extremist and terrorist content, as well as the main extant functions of varying online platforms. Secondly, in light of the diversity of these platforms' functions, it is expected that platforms may use this tool in a variety of ways (as emphasized in the "How to use this tool" section), on the understanding that a generic model like the grid below forms a shared starting point in the detection and moderation of terroristic content. Finally, while using this work product and the subsequent recommendations are not binding on GIFCT members (or any other technology companies), it is intended that this tool will provide a concrete step forward in tackling the scourge of terroristic and violent extremist content.

Context and Assumptions

The GIFCT community understands that the current hash-sharing database is an important but narrow tool to limit extremist and terrorist exploitation of online platforms. It prioritizes the U.N.'s short list of highly lethal and active proscribed terrorist organizations, all of which produce formal propaganda, but does not currently apply to all active terrorist organizations and actors (whether or not they appear on any proscribed list), nor to content produced within less organized violent extremist movements. Further, the hash-sharing database lends itself well to image and video content moderation on platforms designed to share content, but less so (or not at all) on platforms that offer different functionalities also exploited by violent extremist movements.

This IAC working group is aware of and supports GIFCT's call for briefing papers to explore the broadening of the hash-sharing taxonomy, but recognizes that expanding the hash-sharing database itself is not fully sufficient to address the diversity of content-types nor platform-types that comprise the online violent extremism ecosystem. Given this context, our intent is to offer a flexible paradigm to GIFCT member companies and partner organization Tech Against Terrorism for thinking about how to limit extremist exploitation of platforms more broadly as new content and platform-types and new violent extremist perpetrators emerge in the future. An expanded hash-sharing taxonomy should nest within the proposed paradigm without issue, but the proposed paradigm should be useful for categorizing other counter-exploitation behaviors beyond just hash sharing.

For the purposes of this resource, this work product considers violent extremists and terrorists to be those entities that use, threaten, or encourage violence (including against property) in order to harm a perceived enemy and advance an ideological goal. The term "ideological" is understood broadly and used here to include political, religious, economic, or socio-cultural worldviews. The paradigm is primarily based on the behavioral aspects of user-generated content and agnostic as to the specific ideology motivating the creation of that content.

Content in support of violent extremist and terrorist entities may or may not depict violence explicitly, but may support other behaviors of the violent extremist or terrorist entity (such as financing or recruitment) that ultimately supports violence. Therefore, this resource includes content-types including (but not limited to) violence or calls for violence.

This resource is not prescriptive. Different companies will choose differing content moderation strategies across content-types, and potentially by perceived severity of content within a given content-type. Similarly, terrorists and violent extremists can use different platform-types to strategic effect, for operational purposes, or for tactical ends, suggesting that companies may need to prioritize their counter-exploitation efforts appropriately. In addition, not all platform-types will need to address every content-type, either because their functionality does not support that content-type or because the violent extremists and terrorists found on their platform do not engage in the use of certain content-types.

Given that the tech industry is continually bringing new capabilities to the market (platform-types), and users are continually finding new ways to exploit those capabilities and create new content-types, it will be important to update this paradigm regularly. In addition, as terrorist and violent extremist movements evolve and emerge over time, resources of this type need to be revised and consistently checked for accuracy and applicability in order to capture new ways that terrorists and violent extremists exploit digital technology. Likewise, since the primary audience for this work product is GIFCT tech companies and practitioners, this resource has been designed with practical use and moderation consistency in mind.

How to use this resource

By providing a shared paradigm for understanding the landscape of content- and platform-types, including the vulnerabilities of certain platforms to certain kinds of content or violent extremist behaviors, this resource aims to empower GIFCT member companies to better conduct risk assessments, communicate with one another about vulnerabilities and mitigation strategies, move beyond binary content or account takedown decisions to more nuanced moderation strategies, and consider other ways they can mitigate online harms while protecting free speech (such as helping engineers obviate certain platform-specific vulnerabilities in the design process).

This resource is primarily designed for GIFCT companies that host user-generated content rather than for infrastructure providers. The resource equips companies with a greater understanding of the different types of terrorist and violent extremist content online and how these might manifest across different types of platforms. As platforms have different functionalities, design features, and affordances, the prevalence of different content-types on platforms will vary. This resource empowers companies to conduct a risk assessment of content relating to terrorism and violent extremism on their platform(s) and then to set tailored moderation policies and practices according to the type, prevalence, and severity of content. Companies are encouraged to assign differentiated values to content on their platform(s) based on these factors (such as numerical scoring, color coding, or a 'traffic light' risk evaluation system) to enable a more nuanced and prioritized moderation approach that goes beyond binary takedown decisions. It may be useful, for instance, for companies to note the engagement metrics for certain types of content over others, including how widely and quickly content is disseminated, in order to inform appropriate moderation and to ensure terrorist and violent extremist content is not promoted to users.

For further detail on how platform functionality can affect the type of content on a specific online platform, we recommend engaging more closely with partner organizations such as Tech Against Terrorism, which supports companies in improving their understanding of the terrorist and violent extremist threat landscape online.

Classifying terrorist and violent extremist content by platform function

Content	Calls to Action	Ideological / Strategic Content	Material Support	Recruitment, mobilisation and retention
Platform				
Strategic				
Takedown circumvention				
Messaging beacon				
Video streaming				
Operational				
Attack planning				
Automation				
Community maintenance				
Content generation				
Content sharing				
Content storage				
Financing				
Sale of physical goods				
Tactical				
Encrypted communication				
Non-encrypted communication				

Category	Subcategory	Definition	Potentially vunerable platform function examples (illustrative and non-exhaustive)
Strategic			
High impact effects and large audience reach			
Messaging beacon		Use of messaging platform to share outlinks to content stores	Telegram, WhatsApp
Takedown circumvention	Archiving	Archive content stores to record content in case of takedown	Internet Archive
	Parallel content posting	Share content simultaneously across different content stores	MirrorAce, Multifilemirror, MultiUp
Video streaming	Live streaming	Share / record video in real time	Twitch, Facebook, Instagram
Operational			
Moderate impact effects with moderate audience reach and high participation by terrorist actors			
Attack planning	Accomodation	Booking accommodation	AirBnB
	Hostile open-source intelligence	Using online tools to plan an attack	Google Maps, Google Earth
	Travel booking	Arrange travel for meeting, carrying out attack	
	Vehicle hire / usage	Hiring / using vehicles	Uber, Careem
Automation	Bot network tools	Automated tools to create accounts and content	
	Content automation	Automated content posting	Buffer
	Link generation / shorteners	Shortening and obfuscating links	Bit.ly
Community maintenance	Group messaging	Share messages and content with a group	Various
Content generation	Audio sharing	Publishing audio	SoundCloud, Spreaker

Category	Subcategory	Definition	Potentially vunerable platform function examples (illustrative and non-exhaustive)
Content sharing	Audio sharing	Publishing audio	SoundCloud, Spreaker
	Blogs / content management systems	Sharing material via blog	WordPress, Wix
	Book reviews	Publishing book reviews	Good Reads
	Document editing	Sharing editable documents	Google Docs, Office 365
	Image / photo library	Sharing images	Flickr
	Pasting sites	Pasting material anonymously	JustPaste.it
	Video		
	· Partially	Sharing videos	Odysee, DTube
	· Video sharing / archiving	Sharing videos	YouTube, Vimeo, Mega
Content storage	File server		
	· On premise	Sharing folders and files	NextCloud, OwnCloud
	· Cloud	Sharing folders and files	AWS
		Sharing folders and files	IPFS
	File storage	Sharing folders and files	Top4Top, WeTransfer, OneDrive, iCloud, Mail.ru, Zippyshare
	Messaging app	Sharing content within app to other users / via URL	Telegram
	PDF / document storage	Sharing documents	PDFHost, Scribd
Content generation	Deep fake generators	Creating fake videos	
	Gaming	Creating game simulations, mods	Roblox
Financing	Crowd funding / charitable donations	Obtain public funding	GoFundMe, Patreon
	Payment processing	Accepting donations	Crypto exchanges, PayPal
	Selling goods		eBay, Etsy
Sale of physical goods	Books	Sell books / manifestos related to violent ideology	Amazon, Good Reads
	Merchandise sales	Sell goods promoting violent ideology	eBay
Social Media	Alt-tech	So-called alternative tech	Parler, Gab, Bitchute
	Large platforms	Large-scale social media	Facebook, Twitter, Instagram, TikTok, Ok.ru

Category	Subcategory	Definition	Potentially vunerable platform function examples (illustrative and non-exhaustive)
Tactical			
Low impact effects with limited audience and limited range of participation from terrorist actors; however, can be used to facilitate violence or other harms			
Encrypted communication	Closed group / individual messaging	Fully / partially encrypted messaging	Telegram, WhatsApp, Threema, Signal, Element,
	VPN	Secure internet traffic	
	Email		ProtonMail, Mail.ru
	Phone-to-phone messages	Messaging services integrated into phone operating system	iMessage
	Real-time video / calls	Conducting meetings / calls online	Zoom, Teams, Skype
Non-encrypted communication	Gaming messaging	Messaging on gaming platform	Xbox
	Phone-to-phone messages	Messaging services integrated into phone operating system	SMS
	Other messaging platforms		Hoop

Categories and characterization of terrorist and violent extremist content

Subcategories	Description
Calls to Action	
Online material that contains admissions, statements of intent, and aspirational or inspirational statements or content that encourage activity on behalf of or in support of terrorism or violent extremism	
Promoting (non-violent) criminal behaviors in support of terrorism or violent extremism	Online material that encourages non-violent criminal behaviors in support of terrorism or violent extremism (e.g., instructions on the manufacture, acquisition, transfer, or release of prohibited explosives or weapons; the seizure of goods or material to benefit violent extremist groups or organizations, etc.)
Promoting action in support of terrorism or violent extremism	Online material that encourages or promotes activity on behalf of (or in support of) terrorist or violent extremist organizations (e.g., publicity in support of a rally or on behalf of a proscribed organization or violent extremist group)
Promoting propaganda creation	Online material that encourages or facilitates the creation of propaganda in support of terrorism or violent extremism (e.g., visual or graphic templates for branding and distribution of material glorifying violent acts and used as inspirational material, etc.)

Subcategories	Description
Promoting financing	Online material that encourages or facilitates financing for the support of terrorist or violent extremist groups or organizations (e.g., material promoting donations or funding to specific groups or accounts; links to off platform fundraiser)
Promoting criminal activity	Online material that encourages criminal activity in support of terrorism or violent extremism (note that “criminal activity” is subject to varying legal frameworks and jurisdictions)
Promoting individual acts of violence	Online material that encourages or incites individuals to commit acts of violence (e.g., text, visual, or audio material urging specific criminal acts, distribution of terrorist attacks intended to inspire copycats, instruction manuals on how to commit specific offences, etc.)
Promoting group mobilization	Online material that encourages supporters of terrorism or violent extremism to mobilize for a specific purpose or objective (e.g., text, visual, or audio material urging individuals or groups to commit specific activity on behalf of or in support of the group)
Promoting recruitment	Online material encouraging the joining of a terrorist or violent extremist group (e.g., text, audio, or visual material calling on individuals to join the group)
Other	
Ideological / Strategic Content	
Online material that contains narratives that provide engagement with terrorism or violent extremism and/or are intended to advance strategic objectives	
Visual symbols of terrorist or violent extremist movements and organizations	Online material that contains symbols of terrorist or violent extremist movements and organizations (e.g., graphic designed banners or memes, video material, etc.)
Manifestos	Online material of terrorist or violent extremists that is replicated via text, audio, or visual mediums and containing ideological narratives, motives, statements of intent, etc. (e.g., using parts of a religious or ideological text by a terrorist or violent extremist group to justify further violence motives or intents)
Sermons, speeches, or dialogues	Online material of terrorist or violent extremist groups that is replicated via text, audio, or visual mediums in support of ideological or strategic objectives (e.g., quotes of specific sermons used in extremist narratives, etc.)
(Auto-)biographies of extremists for recruitment	Online material that is replicated in full or in part via text, audio, or visual mediums in support of strategic, operational, or tactical objectives (e.g., video autobiographies about an extremist’s experience intended for recruitment purposes)
Key texts	Online material that is replicated in full or in part via text, audio, or visual mediums and relates to specific ideological narratives of a terrorist or extremist group (e.g., key texts used to justify attacks on a specific person or group)
Symbols and coded messages	Online material that contains symbols and coded messages associated with terrorism or violent extremism (e.g., hand gestures signifying white supremacy, etc.)

Subcategories	Description
Propaganda material	Online material that is intended to glorify terrorism or violent extremism, encourage recruitment, radicalize individuals or groups, or otherwise serves some strategic, operational, or tactical objective of such groups (e.g., content glorifying a terrorist act)
Cultural artifacts produced or appropriated for terrorism or violent extremism	Online material such as music, art, or poetry that is used to encourage violence or terrorism towards perceived enemies
Other	
Material Support	
Online material that encourages or is useful to individuals or groups preparing an act of terrorism or extremist violence	
Travel support	Online material that enhances capacity for movement of individuals or transportation of materials (e.g., purchasing travel tickets or accommodation, organizing pick up of individuals or weapons, etc.)
Fundraising	Online material that augments the financial capacity of a terrorist or violent extremist group or individual (e.g., donation campaigns, money transfers, etc.)
Provision of materials	Online material that instructs how to obtain or create materials for terrorist or violent extremist purposes or actions (e.g., recipes to fabricate explosives; material preparations for violent intent or terrorism)
Administrative support	Online material that aids in the management, collection, or distribution for violent extremist or terrorist use (e.g., assistance in purchasing materials for terrorist use)
IT support	Online material that aids in the management or augmentation of computer, technical, artificial intelligence; or any other information technology used by terrorists or violent extremists (e.g., creating web platforms, setup of web infrastructure, managing back end of web platforms; social media spaces, hacking, etc.)
Translation services	Online material or activities aiding the dissemination of terrorist and violent extremist content in different languages
Other material support	
Violent Content	
Attempting to provide a unique definition of online violent content is challenging yet necessary to delineate the boundaries of online violence in protecting freedom of expression. Withstanding multiple interpretations, the following description of violent extremist content is material that consists of the description, reproduction, actual acts, or calls for acts for terrorist or violent extremist purposes, and which are likely to result in physical, sexual, verbal, economic, or psychological harm to individuals or groups	
Depiction or description of a specific act(s) of violence against people	Online material that describes or depicts acts of extremist violence or terrorism (e.g., racial slurs, online harassment, and bullying related to extremist ideologies; trolling; actual acts of physical violence, beating, shooting, etc.)
Promoting or glorifying ideologically-motivated violence against people	Online material that advocates violence against a person or group, and which is justified by a terrorist or extremist ideology (e.g., promoting the elimination of a targeted minority group based on a supremacist or racist ideology)

Subcategories	Description
Depiction of a specific act(s) of violence or vandalism against property	Online material that depicts acts of violence against property when perpetrated on behalf of a terrorist or extremist person or group (e.g., recording of hateful or inciteful slurs on walls of a religious monument)
Promoting or glorifying ideologically-motivated violence against property	Online material that publicizes or advocates violent acts against property for terrorist or violent extremist purposes (e.g., celebrating an act of destruction against a place of worship based on extremist ideology)
Inciting acts of violence against people	Online material that motivates or calls for acts of violence against a person or group based on terrorist or violent extremist ideologies (e.g., online calls for trolling; calls for attacks on women, etc.)
Inciting acts of violence against property	Online material that motivates or calls for acts of violence against property based on terrorist or extremist ideologies (e.g., calls for attacks on a historical, cultural, or religious monument)
Dehumanizing material	Online material that describes, depicts, or encourages the representation of individuals or groups as non-/sub-human or objects for the purpose of supporting, glorifying, or inciting violence
Other	
Recruitment, Mobilization and Retention	
Online material that aims at enhancing the resources and capacities of a group engaged in terrorist or violent extremist activities by augmenting its support, blocking disengagement of supporters, retaining individuals in the movement, or mobilizing them towards action	
Inviting individuals to increased participation in terrorism or violent extremism	Online material aimed at augmenting the support of individuals engaged in terrorist or violent extremist groups, movements, or actions
Event organizing	Online material aimed at organizing events intended to lead to violent action. This category may also include attempts at increasing the number of persons or groups participating in a terrorist or violent extremist event
Threats regarding those who leave terrorist or violent extremist movements	Online material that aims to deter persons from leaving a terrorist or violent extremist group, or taking part in a terrorist or violent extremist act (e.g., threats to persons who wish to disengage from an extremist group; violently targeting former extremists, etc.)
Recruitment into a proscribed organization	Online material that aims at increasing support for a proscribed group (e.g., calls to join a banned terrorist organization). Given that proscription lists evolve and are subject to change, it is vital that due process, human rights concerns, and free speech considerations are included when designating organizations
Recruitment into a terrorist or violent extremist movement or group	Online material aiming at increasing membership in a terrorist or violent extremist group (e.g., recruitment campaigns; online publicity; invitation to gatherings; and information or gaming sessions for purposes of recruitment, etc.)
Other	

Recommendations

The authors of this text propose the following recommendations for GIFCT member companies and partners:

- Develop a more balanced identification of terrorist and violent extremist groups and actors currently operating online. This can be done by developing pre-existing knowledge, frameworks, and resources identifying terroristic Islamist content in the hash-sharing database, as well as developing new knowledge on the specific use of online spaces (such as by right-wing extremist and terrorist groups as well as violent and misogynist groups and actors).
- Expand the hash-sharing database on terrorist and violent extremist content to include not only images (such as memes) and videos, but also URLs (such as PDF files of terrorist manuals). When hash sharing is not the appropriate tool, consider other empowering capabilities that GIFCT can provide member companies (such as natural language processing used in machine learning for identifying texts). Consistently evaluate and improve these capabilities, particularly in order to avoid false positives and human rights violations.
- GIFCT can constructively influence member companies in ensuring that online platforms maintain comprehensive and up-to-date policies.
- Encourage platforms to improve Terms of Service (ToS) by applying the above grid in conjunction with pre-existing policies on hateful conduct and speech, e.g., behavior directly targeting individuals and groups on the basis of race, ethnicity, national origin, disability, religion, sexual orientation, gender, or disease.
- Continually update both this resource as well as the taxonomy of terrorist and violent extremist content, making these evolving resources accessible to GIFCT member companies, ideally through Tech Against Terrorism's Knowledge Sharing Platform.

Indicative Resources

Databases

The Global Terrorism Database, available at link

Center for Research on Extremism (C-REX) RTV (Right-Wing Terrorism and Violence) Dataset, available at link

Tech Against Terrorism, Terrorist Content Analytics Platform (TCAP), available at link

Publications

Alexander, Audrey, and William Braniff. "Marginalizing Violent Extremism Online." Lawfare (January 21, 2018), link

Berger, J.M. Extremism. Cambridge, MA: MIT Press, 2018.

Conway, Maura. "Determining the Role of the Internet in Violent Extremism and Terrorism: Six Suggestions for Progressing Research." Studies in Conflict & Terrorism 40, no. 1 (2017): 77-98. link

Fishman, Brian. "Cross-Roads: Counter-Terrorism and the Internet." Texas National Security Review 2, no. 2 (February 2019), link

LaFree, Gary, et al., "Terrorist use of the Internet." Special issue of the Journal of National Security Law and Policy 9, no. 1 (2017), link

Meserve, Stephen A. and Daniel Pemstein. "Terrorism and Internet Censorship." Journal of Peace Research 57, no. 6 (October 29, 2020): 752-763. link

Tate, Emily B. "'Maybe Someone Dies': The Dilemma of Domestic Terrorism and Internet Edge

Provider Liability." Boston College Law Review 60, no. 6 (2019), link

UNODC. "The use of the Internet for terrorism purposes." 2012, link

Authors

William Braniff

William Braniff is director of the National Consortium for the Study of Terrorism and Responses to Terrorism (START) and a professor of the practice at the University of Maryland. He previously led the practitioner education program at West Point's Combating Terrorism Center, served at the National Nuclear Security Administration, and served in the U.S. Army.

Matthew Feldman

Professor Matthew Feldman is a specialist on fascist ideology and radical-right extremism, and directs of the Centre for Analysis of the Radical Right. He has published and spoken widely on this area for academic as well as general audiences, and has acted as an expert witness in various court cases, parliamentary briefing sessions, and other forums on right-wing extremism.

Eviane Leidig

Dr. Eviane Leidig is a postdoctoral affiliate at the Center for Research on Extremism at the University of Oslo. She is also Head of Policy at the Centre for Analysis of the Radical Right (CARR) and an Associate Fellow at the Global Network on Extremism & Technology (GNET).

Adam Hadley

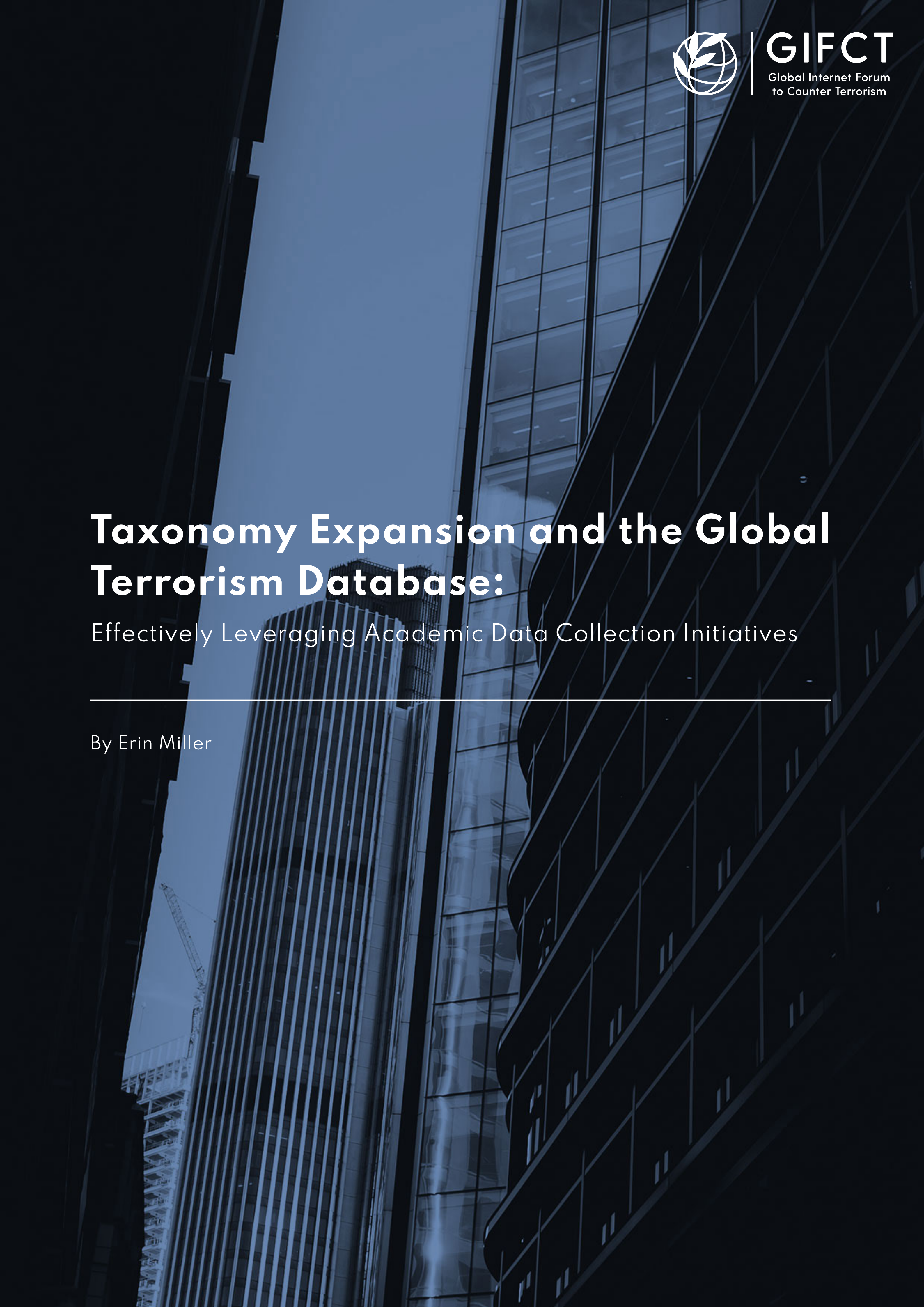
Adam Hadley is the Director of the Tech Against Terrorism Project, a project initiated by the U.N. Counter-Terrorism Directorate (U.N. CTED) in 2016. He is an entrepreneur and data scientist focused on improving data-driven decision making in business and society, and is the CEO of the data science consultancy QuantSpark as well as its not-for-profit arm Online Harms Foundation.

Ghayda Hassan

Professor Ghayda Hassan is a clinical psychologist and professor of clinical psychology at Université du Québec à Montréal and a senior Research Affiliate at Canadian Network for Research on Terrorism, Security and Society (TSAS). She is the director of the Canadian Practitioners' Network for the Prevention of Radicalization and Extremist Violence and co-holder of the UNESCO-PREV Chair on prevention of radicalization and extremist violence.

Ray Serrato

Ray Serrato is a social media analyst and open-source investigator based in Berlin. His research on social media, disinformation, elections, and dangerous speech has been featured in various news outlets. He has worked for the United Nations High Commissioner for Human Rights and for the European Union, the Open Society Foundation, and civil society organizations on projects in countries ranging from Germany and Thailand to Myanmar and Sri Lanka.



GIFCT
Global Internet Forum
to Counter Terrorism

Taxonomy Expansion and the Global Terrorism Database:

Effectively Leveraging Academic Data Collection Initiatives

By Erin Miller

Taxonomy Expansion and the Global Terrorism Database: Effectively Leveraging Academic Data Collection Initiatives

By Erin Miller, PhD

Introduction

One of the key recommendations of the recent GIFCT-sponsored study to evaluate the strengths and limitations of terrorist definitions and designations lists was that "the technology sector and representatives from civil society, academia and government should work together to develop a global, unbiased and real-time database of possible terrorist entities."[1] In comparison to other sources, the authors identified "rigor and objectivity" as an advantage of academic datasets, while drawbacks of academic datasets include poor timeliness and challenges related to sustainability over time.[2] In this briefing paper, I will 1) address each of the issues noted by Meserole and Byman, 2) identify and explore additional opportunities and challenges associated with academic data collection efforts, and 3) articulate specific strategies for effectively and responsibly leveraging academic data collection frameworks to expand the GIFCT taxonomy.

The Global Terrorism Database (GTD) team has developed and maintained rigorous data collection tools and workflows for more than 15 years, balancing artificial intelligence and subject matter expertise to produce consistent, reliable information about terrorist violence around the world.[3] These processes—which are based on a pipeline of real-time, open-source media—include evaluating the validity of source materials, recognizing the potential for and minimizing political influence, and basing inclusion decisions on behavior rather than ideology or third-party designations of terrorism. Focusing on behavior means that any taxonomy expansion strategy that leverages the GTD framework would not be an effort to identify groups or individuals for censorship, but to identify content that promotes violence or recruitment to violence to advance any ideological objective, including emerging threats. While the GTD is an event database, it can be used to compile detailed information about attack locations, perpetrators, targets, tactics, and outcomes. The team recognizes that there is no one-size-fits-all definition of terrorism and strives to capture contextual information that allows users to filter events based on inclusion criteria and make informed decisions about the nature and severity of violence, as well as relevant actors and constructs. Meserole and Byman correctly identified timeliness and

[1] Chris Meserole and Daniel Byman, "Terrorist Definitions and Designations Lists: What Technology Companies Need to Know," Royal United Services Institute for Defence and Security Studies, (2019): 2, link
[2] Meserole and Byman, "Terrorist Definitions," 7-8.
[3] Gary LaFree, Laura Dugan, and Erin Miller, Putting Terrorism in Context: Lessons from the Global Terrorism Database (London: Routledge, 2015); START Global Terrorism Database, "Codebook: Inclusion Criteria and Variables," National Consortium for the Study of Terrorism and Responses to Terrorism, University of Maryland (2019), link

sustainability as critical challenges. The GTD team has specifically worked to address these inter-related challenges on several fronts. I will focus first on issues related to sustainability and then review several strategies for improving timeliness.

Sustainability

The costs associated with maintaining robust, comprehensive data on terrorist attacks around the world are indeed significant—particularly in comparison to most academic research projects in the social sciences, but much less so in comparison to public and private counter-terrorism expenditures around the world. A rough estimate of the annual cost of data collection for the GTD is approximately $2 million, but the actual amount varies depending on a number of factors, including 1) economies of scale between core event data collection and related projects, 2) frequency and timing of data deliveries, 3) project management costs associated with harmonizing multiple diverse sources of funding, and 4) the extent to which the initiative is aimed at maintaining ongoing data collection workflows or more aggressively working to improve both data and systems.[4] Specific costs associated with the maintenance of the GTD include subscription access to multiple robust and comprehensive feeds of aggregated news articles, a software architect to provide technical expertise and infrastructure for automating and maintaining data pipelines both for source document processing and data publication, ten subject matter experts for terrorist attack identification and detailed coding, an experiential education initiative to train and mentor students assisting with data collection, project management and administration, and data dissemination. These costs represent a fairly moderate approach for maintenance and do not include, for example, resources for research and development on artificial intelligence applications to improve the efficiency of data collection workflows or investments in improving multi-lingual capabilities of data collection tools and personnel to improve the comprehensiveness and representativeness of the resulting data.

Sustaining a resource like the GTD requires stable and consistent funding. Since 2002, work on the GTD has been generously sponsored at various times by the U.S. National Institute of Justice, the U.S. Department of Homeland Security, the U.S. Department of State, the U.S. Department of Defense, the German Federal Foreign Office, and the U.K. Foreign, Commonwealth, and Development Office. During recent lengthy lapses in government funding, the University of Maryland has used reserve funds to partially bridge gaps.

A core challenge to sustainability is the paradoxical situation that, as a public good, the

[4] Meserole and Byman note that "The winning bid for the most recent Department of Homeland Security grant for a global terrorism database had a proposed budget of over $10 million. See U.S. Government Accountability Office, 'University of Maryland', B-416682, 24 October 2018" ("Terrorist Definitions," 6). To clarify, the grant in question was administered by the US Department of State rather than the U.S. Department of Homeland Security, and the $10 million budget covered five years of performance.

GTD is relied upon by an incredibly broad and diverse array of users, including government analysts and policymakers, researchers and students, and private companies across numerous industries, yet developing strategies to promote cost-sharing among these users drastically increases the costs of maintaining the resource. Establishing one organization or agency as a centralized source of funding creates significant vulnerabilities and a single point of failure when that source is cut. It is also somewhat counterintuitive to the idea that cost-sharing among those who rely on the data is an inherently fairer arrangement. However, the administrative costs associated with managing technical requirements of various agencies are considerable, and in our experience with the GTD, this strategy has not eliminated the risk of gaps in funding. Likewise, attempting to engage the private sector in sharing the costs of data collection through paid licensing also incurs significant new costs associated with sales, marketing, customer service, advanced access control, and development of analytical tools. A paid commercial licensing strategy also faces major obstacles to success for a product that—as a public good aimed at providing transparent data on a contentious subject in order to promote understanding and security—must remain freely available to the research community and general public.

This collective action dilemma, known as “the free rider problem,” is not unique to the GTD. The Stanford Encyclopedia of Philosophy (SEP)—itself a fragile public good with a free rider problem—links the concept to Plato, Adam Smith, David Hume, John Stuart Mill, and Vilfredo Pareto.[5] In explaining their own decision to not pursue commercial licensing for the SEP, its editors concluded that such action would dramatically increase costs, reduce the impact of the product due to heavily restricted access, and ultimately would not guarantee long-term sustainability. They write:

> The above results all combine in pernicious ways, with the result being that a subscription-based funding model would lead the SEP project towards a situation where it loses it [sic] focus and character as a project developed, administered and maintained by academics. Not only would the SEP reach a tiny fraction of the audience it once reached, but it might be forced to scramble each year to make ends meet, distracting its central staff from the academic mission of enhancing the encyclopedia’s content and technological underpinnings.[6]

Developing solutions to promote data sustainability is a complex challenge shared by many. The Open Data Institute conducted a multi-year research and development program to identify strategies and build tools to design sustainable data institutions.[7] Their analysis of the business and revenue models of existing data institutions indicates that it is common for data institutions to use mixed funding streams and develop contingency plans and that effective governance and community support are critical to ensure that the

[5] Russell Hardin and Garrett Cullity, “The Free Rider Problem,” in The Stanford Encyclopedia of Philosophy, ed. Edward N. Zalta, Winter 2020 Edition, link
[6] John Perry, Edward N. Zalta, Uri Nodelman, and Colin Allen, “The Problems with a Traditional Funding Model,” Center for the Study of Language and Information, Stanford University, (n.d.).link
[7] ODI, “Designing Sustainable Data Institutions,” Open Data Institute, (2020), link

institution continues to serve the needs of the community rather than particular interest groups. Likewise, ODI's study identified myriad examples of problematic tensions between revenue models and organizational goals and values. For example, they note that "the need to maintain independence around the purpose of a data institution and the goals of its community can create tension when using revenue from funders or investors that have their own goals or priorities."[8]

Faced with the need to balance sustainability, resilience, efficiency, and transparency, START has been pursuing an "all of the above" approach, simultaneously advancing multiple strategies for securing consistent funding for the GTD. Which one (or combination of these) is most likely to ultimately be effective remains to be seen. In fact, at the time of this writing, the GTD project is experiencing another gap in funding, albeit one we hope to be resolved in a matter of weeks. In the meantime, we continue to work to develop mechanisms for governments to pool resources and share costs and to seek out meaningful collaboration with private sector organizations representing broad industry segments to streamline investments in security. What is clear is that continuing to build robust partnerships with institutions that share our values will be essential to our success.

Timeliness

Insufficient timeliness of academic datasets like the GTD is the second limitation raised by Meserole and Byman. Although the historical data in the GTD is certainly relevant to the issue of taxonomy expansion, as I discuss below, the GTD's current one-year lag behind real time makes it ill-suited for rapidly identifying emerging organizations, movements, and related content that may be appropriate for inclusion in the GIFCT taxonomy. Since the GTD project began in 2002 with an initiative to digitize the event data that ended in 1997, "catching up" has been a central feature of our efforts.[9] Clearly, the issue of timeliness is closely linked to available resources and overall sustainability. However, even despite sustainability challenges, the GTD team has endeavored to continually improve timeliness and has enjoyed modest success. Currently, in mid-2021, the GTD team is collecting data on events that took place in mid-2020. To date we have pursued three general strategies to help minimize the lag behind real time: 1) Personnel Increases, 2) Artificial Intelligence and Automation, and 3) Prioritization of Preliminary Data.

Personnel Increases

Increasing the personnel working on data collection as resources allow is the most straightforward "brute force" strategy. It does not involve any adjustments to workflows—only training more people to do the work of reviewing open-source news articles and systematically identifying events that meet the GTD's definition of a terrorist attack. The key to this strategy is recruiting enough staff to overcome the particular vagaries of the academic calendar. The GTD project is, after all, an academic research project with

[8] ODI, "Designing Sustainable Data Institutions," 34.

[9] LaFree, Dugan, and Miller, Putting Terrorism in Context, 2015.

scholarly objectives, so the research team's ability to continually track terrorist attacks is impacted not only by the usual holidays and leave time, but gaps between semesters, periods of training for new internship cohorts, midterms weeks, final presentation weeks, and so on. Reducing the lag time created by these gaps and gaps in funding requires hiring enough research personnel to effectively maintain the pace of data collection. We have, during times of relatively reliable funding, been quite effective at this. For example, in 2019 the GTD team increased the size of the research team by 2-3 FTE and correspondingly reduced the lag time for fully completed data to three months and preliminary identification of new events to a few days. Unfortunately, gains achieved through personnel increases are easily eliminated when lapses in funding lead to partial or complete work stoppages.

Artificial Intelligence and Automation

The GTD team's data collection methodology balances artificial intelligence and automation against subject matter expertise to maximize the strengths and minimize the limitations of each. The workflow starts with subscription access to news feeds that aggregate more than two million articles published daily, and the task is to find those "needles in the haystack" that describe terrorist attacks. Automation is good for processing large numbers of documents very quickly, but artificial intelligence is not sufficiently effective at synthesizing unstructured text that often contains conflicting or vague information. Subject matter experts are far more effective at navigating conflicting or vague information in unstructured text, but they lack the capacity to read two million articles per day. Due to the realities of natural language, subject matter experts must cast a very wide net to isolate news articles that describe violent terrorist attacks—a simple keyword search for "terrorism" is not nearly sufficient. The GTD team uses a number of automated strategies to isolate and organize the news articles most likely to contain relevant information about terrorist attacks. Researchers then review these articles and manually extract records of unique events that qualify for inclusion in the GTD.

We are always interested in opportunities to adjust this balance between artificial and human intelligence and lean a bit more on automation where it is possible to do so and improve efficiency without compromising the quality of the data. For example, at times when individual terrorist attacks generate an extraordinary amount of media coverage, it is not an effective use of time for researchers to review and discard the hundreds or thousands of repetitive news articles that remain even after our automated de-duplication process. To minimize this inefficiency, we began routinely reviewing the news articles to identify days where the number of articles published is more than one standard deviation higher than the daily average for a given month. The team determines if these outlier dates are driven by media coverage of a particular event. If so, an analyst will seek out a selection of news articles that sufficiently document the event, and the GTD's software developer will automatically remove the remaining articles about that event. The first month the GTD team used this strategy was April 2013 to process the high volume of news articles published about the Boston Marathon bombing and the search for assailants that followed. We began checking for such outlier events each month as standard practice after the November 2015 attacks in Paris and the subsequent search for assailants.

The GTD team has made other advances leveraging automation to improve the efficiency of data collection workflows and help improve the timeliness of the data. We also have planned strategies for additional, more sophisticated applications of artificial intelligence, which the team could incorporate into existing data collection methods. Conducting the research and development to implement additional automation strategies while maintaining the accuracy of the data would have a number of potential benefits, including enabling more comprehensive and rapid detection of emerging groups, movements, and related content that may be relevant for inclusion in the expanded taxonomy.

Prioritization of Preliminary Data

The time required to produce the GTD is largely a result of the extensive level of detail included in the database. The GTD includes more than 100 variables documenting the location, perpetrators, targets, tactics, and outcomes of each attack. This richness of information is undoubtedly a key strength of the GTD, facilitating multidimensional analysis of patterns of terrorism. However, to arrive at this final product the data collection workflow takes place in multiple stages. After the automated pre-processing of source documents is completed, the research team's first task is to review the news articles most likely to contain relevant information and create preliminary records of individual attacks (a process called "triaging"). These preliminary records contain very little structured information—just a brief description of the attack that allows the team to differentiate it from other events. Once the researchers have completed the process of triaging the news articles published in a given month and creating the initial records of attacks in the database, the set of events documented for that month moves to the next stage of data collection, where coders proceed to review the source documents for each record and complete coding for the full complement of structured variables in the GTD codebook. The first step of identifying terrorist attacks, triaging, typically takes place several months in advance of the full coding process and presents an important opportunity for capturing preliminary data.

In April 2021, the GTD team began a pilot project designed to evaluate the validity of real-time data and feasibility of real-time workflows. Temporarily pausing ongoing data collection for events that took place in 2020 to conduct real-time data collection for the month of April 2021, we added several simplified versions of key variables to the triaging process. GTD researchers triaged news articles within a few days after their publication date, creating initial records of attacks that took place based only on the information available in the immediate aftermath of the attack.

To complete the pilot project, the team will ultimately finish the April 2021 data using the full GTD coding process, multiple stages of quality control, and the benefit of hindsight from the information published in May and June 2021. We will conduct a systematic comparison of the data recorded in the immediate aftermath of the attack in April and the data recorded in the weeks and months that followed. In addition to generating valuable insights about the ways in which real-time data collection strategies both help and hinder efficiency, we will identify the strengths and limitations of early reporting and offer users guidance on the merits of preliminary data. Our questions include to what extent does information on inclusion criteria, attack location, perpetrators, targets, tactics, and outcomes evolve as new reporting becomes available, and are certain variables based on initial reports generally reliable while others are particularly unreliable? With this information, we can consider the feasibility and usefulness of disseminating limited preliminary data for use in certain applications that stand to benefit from more immediate access to data that may have tolerable sources of error.

Opportunities

My goal in delving into the issues of sustainability and timeliness is to provide readers with a greater appreciation for the nuances of these challenges and the ongoing efforts already underway to address them. Specifically, I want to make it clear that these limitations are eminently solvable given effective strategic partnerships and continued innovation. Indeed, there are many opportunities to leverage the objectivity and comprehensiveness of the GTD to help inform the GIFCT taxonomy. There are also many reasons that the GTD and the data collection workflows established by the GTD research team present useful foundations to build on to create a taxonomy expansion strategy that has the flexibility and rigor required to be effective across diverse applications.

The GTD is global in scope. All decisions about inclusion in the database are made by an independent research team extensively trained on the fundamentals of social science research methodology to maximize the accuracy of the data and acknowledge potential sources of bias. The information in the GTD is derived from open-source journalism to provide transparency about the decision-making process, and measures of uncertainty are built into the structured data. The GTD research team makes decisions about whether or not to include an attack in the database based on the characteristics of the event and whether or not it meets the established thresholds for terrorist violence. These decisions are not based on a priori designations of individuals or groups as “terrorists” issued by governments or other entities. This focus on all ideologically-motivated violent behavior rather than particular ideas or beliefs is consistent with the needs highlighted by GIFCT: both to identify emerging threats that may not fit neatly into existing frameworks as well as the need to protect human rights and non-violent expression. Because the GTD is an event-level dataset, it allows users to assess the severity of a particular attack or a series of attacks attributed to a particular individual or organization. It also supports taxonomy expansion that goes beyond focusing on particular organizations and includes other units of analysis such as individuals, movements, events, tactics, documents, and symbols directly linked to acts of violence.

With this in mind, there are several key examples of opportunities to use the data and tools developed by the GTD team to inform taxonomy expansion. Note that I am not advocating for a specific substantive taxonomy but illustrating general frameworks and tools that can be useful resources. First, it is helpful to consider the following three categories of content published on social media platforms, distinguished by their temporal orientation:

1. Content published in real time referencing events and actors emerging in real time. For example, as a terrorist attack is taking place or in the immediate aftermath, social media users may publish content addressing the attack or the assailant.
2. Content published in real time referencing historically relevant events and actors. For example, on the upcoming 20th anniversary of the September 11th attacks in the United States, social media users may publish content addressing the September 11th attacks or assailants.
3. Content published historically referencing historically relevant events and actors. For example, at the time of an attack that took place a decade ago, social media users may have published content addressing the attack as it happened, or addressing events that took place even earlier.

Insofar as any of these three types of content may be problematic and potentially suitable for review or removal by content moderators, the information in the GTD can help establish a framework for identifying it. The appropriate strategy varies somewhat depending on the type of content in question.

The most straightforward approach to leveraging the GTD would address the second and third types of content because references to historical actors and events are not impacted by the timeliness (or lack thereof) of GTD production. The GTD's records back to 1970 could be used to identify historical actors—including individuals, groups, and movements that engaged in terrorist violence—and significant attacks likely to persistently provide fodder for those intent on promoting and glorifying terrorist violence online. Consider the 2015 shooting at the Charlie Hebdo offices in Paris, the 1995 bombing at the Alfred P. Murrah Federal Building in Oklahoma City, the 1985 bombing of Air India Flight 182 en route from Montreal to London, or the 1975 hostage-barricade attack targeting OPEC leaders in Vienna. The decision on whether favorable references to these attacks intended to praise them or spur others to action is suitable for removal remains the responsibility of individual companies. However, the GTD could be used as is to help identify references to high-profile terrorist attacks and the individuals and groups responsible for them, provide decision makers with useful context, and supply independent, transparent justification for review and possible removal.

A more sophisticated approach would also address the first type of content—real-time commentary about actors and events as they emerge. Even the most egregious examples of this, such as the livestreamed 2019 attack targeting mosques in Christchurch, New Zealand, are difficult to detect and address rapidly using existing tools.[10] The challenge is

[10] New Zealand Ministry of Foreign Affairs and Trade, "Christchurch Call to Eliminate Terrorist and Violent Extremist Online Content," originally posted 2019, link

even greater for the hundreds of terrorist attacks that take place around the world each month. Terrorist attacks do not all generate the same amount of attention on social media platforms, nor do they have the same likelihood of being used as a call to action online. But, if an objective of the GIFCT taxonomy expansion initiative is to more comprehensively and uniformly identify problematic content, one way to accomplish that is to build on the data collection processes and tools that the GTD team already uses to curate source documents and extract information on violent attacks.

The April 2021 pilot project described above established a proof-of-concept for the real-time triaging of GTD source documents and provided valuable insight into issues of efficiency and adaptations to the triaging process designed to capture preliminary data. The fact that the research team is already triaging source documents and recording preliminary details about terrorist attacks presents a useful opportunity to add questions about the potential notoriety of attacks as they happen. Like the existing GTD collection strategies, this could involve both automated techniques and subject matter experts focused on early detection. Automated techniques could be used to model the likelihood of a terrorist attack or an attacker generating praise or calls to action on social media based on the characteristics of the attack or even the characteristics of the conventional media coverage of the attack. The latter technique capitalizes on the triagers' extensive experience reading about terrorist attacks to flag events that, in their informed estimation, are likely to function as violent propaganda. Relevant indicators may include the emergence of previously unidentified perpetrator groups, innovative or extraordinarily provocative tactics and attention-seeking behavior, or attack motivations known to be frequently promoted on social media platforms.

Implementation

Having spent more than 15 years developing tools and workflows to support systematic collection of data on terrorism, I would be remiss if I did not acknowledge the profound implementation challenges associated with taxonomy expansion. Transitioning from the relatively narrow focus of the current taxonomy to a framework that more uniformly and comprehensively captures terrorist threats is far easier said than done. As difficult as it is to devise a strategy for conceptually expanding the taxonomy, it is much harder to apply an expanded taxonomy to content in the wild. Those working on the conceptual question should take this into account. For example, introducing the inclusion of terrorist groups that are lesser known and less effective at branding and public relations than al Qaida and Islamic State means creating tremendous ambiguity about whether a piece of content can be reliably linked to an entry in the taxonomy. Adding more key events or high-profile assailants to the taxonomy further increases the difficulty of determining what material published on social media platforms is a match. Finally, to the extent that diversifying the types of content under review to include more than hashed images and videos is a priority, the implementation difficulties increase exponentially.

Because using and contributing to the hash-sharing database is voluntary, it is critical to maximize simplicity as much as possible in order to promote participation. This is necessary to reduce the technical burden on companies integrating the hash-sharing database API into their content moderation tools, as well as to minimize decision fatigue among the people adding entries to the hash-sharing database. The GTD includes more than 3,600 organizations and movements responsible for carrying out terrorist attacks worldwide since 1970. Each year, the GTD team identifies around 50 to 100 new perpetrator groups. The details in the GTD about the attack patterns of these groups, including their tactics, targets, lethality, and regions of activity can help objectively prioritize which of them pose a significant threat via social media platforms. This would allow reviewers to focus on the worst of the worst, and optimize the signal to noise ratio as much as possible.

In addition to promoting simplicity, it may be useful to leverage subject matter experts familiar with the GTD and its sources to produce supporting resources such as reference materials, training materials, and technical guidance crafted to facilitate decision making. Such materials could be designed to simplify and distill key information, highlight significant dates like anniversaries, clarify potential points of confusion and ambiguity, and could be expanded and updated as new information becomes available.

Conclusion

The challenges associated with the sustainability and timeliness of the GTD are significant but are not insurmountable. Through our ongoing efforts to address these challenges by building effective public and private sector partnerships while innovating to improve efficiency, the GTD team has developed a data collection platform that is uniquely suited to provide a foundation for novel strategies to support GIFCT in the shared mission to counter terrorism online. The GTD provides robust global data covering five decades; independent and transparent documentation that allows third parties to review decisions; a focus on violent actions rather than identity groups or political designations; detailed attack characteristics that can be used to generate severity profiles of terrorist attacks and actors; and the research team's substantive expertise—more than 50 years of combined experience working on the project team to produce high-quality structured data on terrorism. Making strategic investments in existing academic data collection infrastructure and subject matter expertise is an important part of a comprehensive strategy to produce a taxonomy that more uniformly reflects the evolving realities of terrorist violence.

References

Hardin, Russell and Garrett Cullity. "The Free Rider Problem." In The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Winter 2020 Edition (Stanford University). link

LaFree, Gary, Laura Dugan, and Erin Miller. Putting Terrorism in Context: Lessons from the Global Terrorism Database (London: Routledge, 2015).

Meserole, Chris and Daniel Byman. "Terrorist Definitions and Designations Lists: What Technology Companies Need to Know." Royal United Services Institute for Defence and Security Studies. (2019), link
The New Zealand Ministry of Foreign Affairs and Trade. "Christchurch Call to Eliminate Terrorist and Violent Extremist Online Content." Originally posted 2019. link

ODI. "Designing Sustainable Data Institutions." Open Data Institute. (2020). link

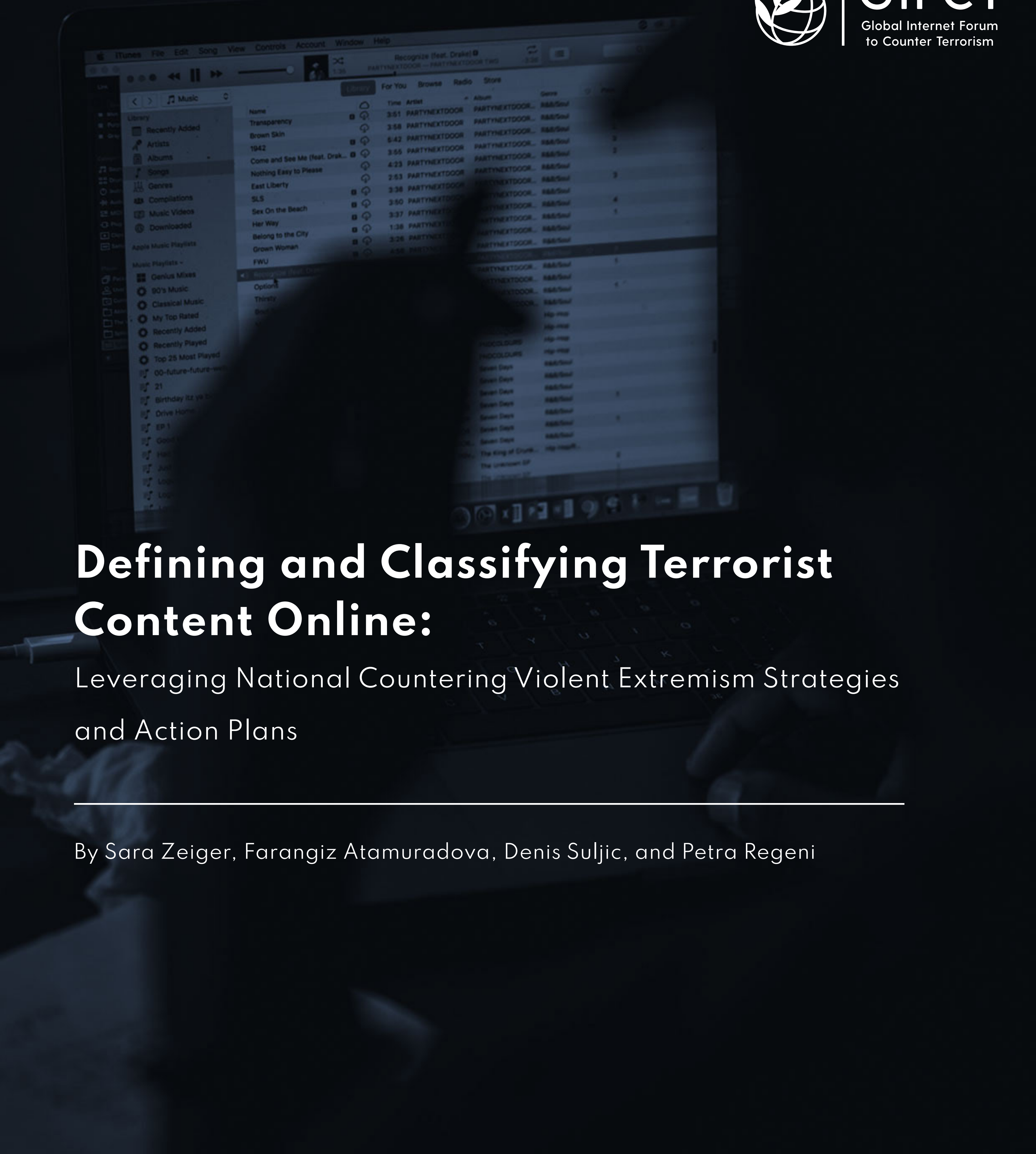
Perry, John, Edward N. Zalta, Uri Nodelman, and Colin Allen. "The Problems with a Traditional Funding Model." Center for the Study of Language and Information, Stanford University. (n.d.). link

START Global Terrorism Database. "Codebook: Inclusion Criteria and Variables." National Consortium for the Study of Terrorism and Responses to Terrorism, University of Maryland. (2019). link

Biography

Erin Miller is an Assistant Research Scientist at the University of Maryland and Principal Investigator for the Global Terrorism Database (GTD) and related research projects. Miller earned a BA in Sociology from the University of Pennsylvania and an MA and PhD in Criminology and Criminal Justice from the University of Maryland. She has been part of the GTD team since 2004, and her roles have included improving the consistency of the data and adding key variables to the database, developing efficient and effective data collection strategies, workflows, and training, and producing accessible analysis that provides context for current events in terrorism and counter-terrorism. She frequently consults with users of the database, including researchers, policy makers, analysts, journalists, and students. For six years, Miller authored the Statistical Annex for the U.S. State Department's Country Reports on Terrorism (2012—2017). Her research investigates patterns of decline among terrorist organizations and movements worldwide, using innovative statistical analysis of data from the GTD. She has taught statistics courses at the University of Maryland and delivered invited lectures on the GTD and the implications of research methodology for terrorism research.

GIFCT
Global Internet Forum
to Counter Terrorism



Defining and Classifying Terrorist Content Online:

Leveraging National Countering Violent Extremism Strategies and Action Plans

By Sara Zeiger, Farangiz Atamuradova, Denis Suljic, and Petra Regeni

Defining and Classifying Terrorist Content Online: Leveraging National Countering Violent Extremism Strategies and Action Plans

By Sara Zeiger, Farangiz Atamuradova, Denis Suljic, and Petra Regeni

Introduction

Social media and digital platforms are important tools for spreading information, communicating across borders in a global world, doing business, and keeping in touch with family and friends. Like the rest of society, terrorist groups and dangerous actors leverage social media and digital platforms to spread their narratives and recruit others to their cause in the digital space. At the same time, digital platforms are also dedicated to preventing terrorists from using their platforms to cause harm. However, the definition of what constitutes "terrorism" is not always well-defined globally or locally. There is no universal definition of "terrorism" or related terms such as "radicalization," "extremism," and "violent extremism." As a result, technology companies need to navigate a complex collection of policy and legislative frameworks to set platform standards and community guidelines around terrorism and violent extremism. They must apply different definitions and employ various list-based approaches to define the parameters of their content moderation policies, community standards, and guidelines.

One source of information about definitions of "terrorism" and other related terms are the national strategies and action plans of the various countries that have been developed in line with the UN Secretary General's Plan of Action for Preventing Violent Extremism.[11] While some countries had already developed documents to support the implementation of programs for preventing and countering violent extremism (P/CVE), the Plan of Action by the Secretary-General also mandated that all Member States have unique plans of action. Subsequently, in the past six years, many countries have developed national action plans (NAPs) or strategies, and many more are in the final stages of approving their new plans.

The research in this paper builds upon the existing NAPs across 35 countries and the European Union[12] and is based on the experience of Hedayah,[13] which has been working with governments to develop and implement national strategies and action plans for P/CVE since 2014. In combination with national legislation on counter-terrorism, NAPs and national strategies often set out the country's definition and approaches to terrorism, violent extremism, extremism, and radicalization both online and offline. Hedayah's work on this subject has granted it privileged access to the conversations, discussions, and

[11] United Nations, Plan of Action to Prevent Violent Extremism, Report to the Secretary General, 2015, link
[12] The list of countries and corresponding documents consulted in this study can be found in Annex A of this report.
[13] Hedayah is the premier international organization dedicated to using its expertise and experiences to CVE in all of its forms and manifestations through dialogue, communications, capacity building programs, research, and analysis.

disagreements on how this work is to be defined in a variety of contexts such as Europe, Central Asia, Southeast Asia, Africa, and the Balkans.[14]

By systematically classifying and coding the definitions of key terms from these documents, this paper identifies common patterns and trends, and analyzes the strengths and weaknesses of different aspects of these definitions. Importantly, this research also identified and coded the key passages that make reference to online engagement, social media companies, counter messaging, and the role of the technology sector. Notably, the dataset in this study is not comprehensive, but a list of the sources, along with several reference guides, are contained at the end of this document (Annex A).

This paper answers the following research questions:

1. How do different countries define terrorism, violent extremism, extremism, and radicalization according to their NAPs and strategies?
2. What is the role of the online space in P/CVE efforts according to the NAPs and strategies?
3. What is the role of technology companies in defining terrorism and efforts to counter it?

The research in this paper is beneficial for several reasons. First, technology companies at a minimum need to be compliant with local laws surrounding the prevention of terrorism and removal of terrorist content from their platforms. However, ideally, company policies also contribute to a broader goal of reducing violent extremism in the countries of operation. In this regard, aligning with national strategies for counter-terrorism (CT) and CVE, NAPs for CT/CVE can help technology companies secure a strong relationship with the governments of the countries in which they operate, and establish a strong partnership between the public sector and the technology sector (provided that the NAPs and strategies also abide by principles of basic human rights and fundamental freedoms mandated by international laws). It should be noted that while this paper does not offer insight specifically on national legislation, nor should it be misperceived as legal advice, the results may provide insight into how to align company policies with national ones. Still, technology companies should also continue to consult official legislation to clarify what countries legally define as terrorism or other terminology related to this subject.

Second, identifying common themes and approaches of governments in their definitions can provide insight for the technology sector for defining terrorism more broadly in community standards within and across platforms, and making adjustments regionally or at the country level. That is, awareness of these trends around definitions can help

[14] In some cases, Hedayah has advised on draft NAPs in their original language and English before these documents are made public. Hedayah takes a localized approach to its work, adapting the definition used in each training to reflect the national laws, strategies, and policies applicable to each country or region.

technology companies tailor the implementation of their company policies to the local context, but also set broad enough policies that encompass many varied aspects of what constitutes terrorism.

Third, this approach also identifies gaps where definitions are not clearly articulated, which is an opportunity for the technology sector and the government sector to collaborate on developing working definitions, and what those definitions mean in practice.

Methodology

The methodology utilized in this paper includes quantitative data collection and qualitative coding of the definition of four key terms (terrorism, violent extremism, extremism, and radicalization) and two key themes (countering violent extremism and internet and social media). The definitions are derived from NAPs and strategies from a pre-selected set of 36 entities that were chosen reflecting a variety of regions and contexts.[15] In light of some missing definitions (namely "terrorism"), the criminal code or national CT legislation was utilized to fill the gaps.

The coding scheme for analyzing the definitions of three terms – "terrorism," "violent extremism," and "extremism" – was identical. The definitions of these three terms were evaluated using the same 16 codes:

- **a. "definition exists":** whether the definition of the term exists in a legal document, national strategy, and/or plan;
- **b. "criminal act":** a clear indication of the term entailing criminal activity, which was assessed by explicit categorization of the act or any other similar indication that clearly assigns the action as a criminal one;
- **c. "threat, incitement, and/or intimidation to violence":** whether the term involves threat, incitement, and/or intimidation to violence;
- **d. "act of violence":** the explicit mention of violent acts (including murder and arson);
- **e. "against public, civil society, and/or civilians":** an action directed against innocent civilians not affiliated with formal institutions and/or the general public;
- **f. "against infrastructure":** physical and non-physical critical infrastructure (including property, cyber/IT, water resources);
- **g. "against political structure":** the government, political institutions, and/or the democratic system as a whole;
- **h. "against economic structure":** banks, financial institutions, and/or the general economy;
- **i. "against social cohesion":** social and cultural values of the community and/or

[15] Including one regional organization (the European Union). We will use the term "countries" to broadly refer to all these entities.

state;

j. **"against constitutional values":** democratic values and governance structure of the country;

k. **"against national identity":** explicit mention of national identity or identity of the country;

l. **"radicalization":** explicit mention of radicalization;

m. **"recruitment":** explicit mention of recruitment;

n. **"support for political causes":** specifying the furtherance of political objectives/causes;

o. **"support for religious causes":** specifying the furtherance of religious objectives/causes; and,

p. **"support for ideological causes":** specifying the furtherance of ideological objectives/causes.

The coding scheme for "radicalization," "countering violent extremism," and "social media" were all different. The codes used and elaborated on in the analysis below targeted the distinct features, keywords, and significant trends unique to each definition/theme.

Finally, since some countries having multiple NAPs and/or general security strategy plans, in addition to legislative documents, the codes for each country under each term/theme were consolidated and merged into one coded line per country. This gave a more comprehensive overview of the codes (present or absent) within each country's overall framework where multiple national documents referenced the terms and themes.

Results & Discussion

Terrorism

All countries were individually analyzed for 16 characteristics of terrorism and coded according to whether one or more of the documents mentioned that trait. As illustrated in Figure 1, every country had an existing definition of terrorism. However, the other 15 codes demonstrate the evident differences between what variables are frequently included and of seemingly higher priority to countries globally.

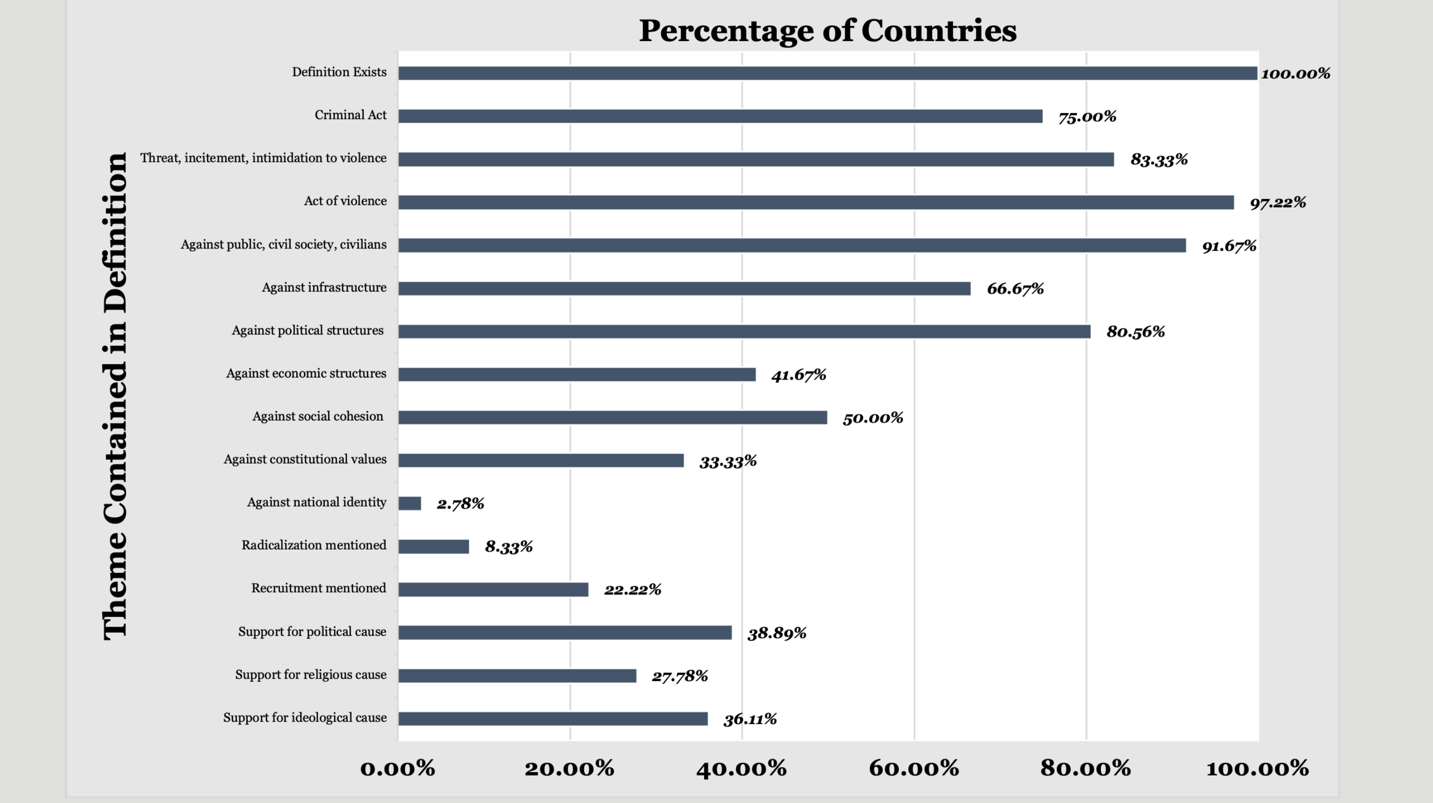


Figure 1. Definitions of Terrorism.

Out of the 15 codes, five are noticeably the most significant traits referred to in definitions of terrorism. The most common code for terrorism was **"acts of violence,"** which was coded in the definitions of 35 countries (97.22%). The code explicitly noted the mention of violence or acts which are characterized are violent. For instance, the **Criminal Code of Turkmenistan** designated "[t]he commission of an explosion, arson or other actions" as acts of violence constituting terrorism."[16] Other countries either follow the format of Turkmenistan, in defining exact acts of violence, or more generally correlate violence with terrorism – for example, Croatia's **National Strategy for Prevention and Suppression of Terrorism** refers to terrorism as "Any act of a criminal nature accompanied by an act of violence."[17] Related to this, the code for **"threat, incitement or intimation to violence"** was

[16] Turkmenistan, Ministry of Adalat of Turkmenistan, Criminal Code of Turkmenistan, 2010, link, (Section 12, Chapter 29, article 271).

[17] Croatia, Government of the Republic of Croatia, National Strategy for Prevention and Suppression of Terrorism, 2015, link, (in "I. The Response of the Republic of Croatia to the Threat of Terrorism" section, under bullet point 3).

featured in 30 countries (83.33%). While some countries mention the threat of violence as an integral aspect of terrorism, others specifically isolate and mention it in subsections – such as the **Criminal Code of the Kingdom of Netherlands**, which distinguishes Article 83 as "Terrorist offences" followed by Article 83a as "terrorist intent."[18] Interestingly, five of the countries coded for acts of violence do not mention the threat or incitement or intimation to violence in their definition of terrorism.

The second most frequent variable was terrorism as an act **"against public, civil society, or civilians"** which was coded in the definitions of 33 countries (94.44%). Following close behind was the code for terrorism directed **"against political structures"** – mentioned in the definition of 29 countries (80.56%). Unsurprisingly, the two most concerning targets of terrorism for countries are civilians and political institutions. Given the fact that NAPs prioritize the lives of their citizens and civil society (often including foreign innocent civilians) and the political structures of government sustaining democracy and social order, it is logical for definitions of terrorism to focus on these elements. All 29 countries that mention attacks against some form of political structures in their definition of terrorism also mention attacks against civilians and civil society – while four countries only mention the latter and omit the former.

The fifth most coded variable across definitions of terrorism was **"criminal act."** This entailed that NAPs and/or legal documents explicitly regard terrorism as a criminal act according to national law. 27 countries (75%) specifically included the criminality of terrorist acts in accordance with laws. Given the fact that prosecuting terrorism has been increasingly prioritized by states (though consequentially more difficult to legally substantiate), the 75% mention rate is seemingly low.

The above five most statistically significant codes were all featured in 21 countries (58.33%) definitions of terrorism, with an additional seven countries (19.44%) missing one of the five codes and an additional six countries (16.67%) missing two of the five codes. Geographically speaking, no striking patterns were evident between countries regionally – with the exception of Central/South Asia, where Kazakhstan, Kyrgyzstan, Pakistan, Turkmenistan and Uzbekistan mentioning all five significant codes, and Tajikistan mentioning all but criminal act. Notably, this means that Central and South Asia have quite robust definitions of terrorism compared to some other regions. However, this may be reflective of the fact that these countries have worked with international organizations such as the UN Development Programme (UNDP), the Organization for the Security and Co-operation in Europe (OSCE), and Hedayah to extensively apply good practices and lessons learned from previously-conceptualized definitions of terrorism in ways that other countries (such as those in Europe) have not.

When it came to defining targets of terrorism, 24 countries (66.67%) mentioned **"against infrastructure,"** 18 countries (50%) mentioned **"against social cohesion,"** and 15 countries

[18] Netherlands, Government of the Kingdom of Netherlands, Criminal Code of the Kingdom of Netherlands, 2012, link, 68.

(41.67%) mentioned **"against economic structures."** The code for infrastructure could mean a vast variety of targets such as the ones specified in the Criminal Code of the Czech Republic, Section 311(1)(c): "Public facility, transportation or communication system including an information system."[19] Although these codes were less prevalent than those discussed previously, several countries gave concise definitions including the above. A comprehensive and well-articulated example from the Philippines defined terrorism as an act to "Intimidate the public and destroy or destabilise the fundamental political, constitutional, economic and/or social structures of a country by causing death or injury to any person, destroying property and critical infrastructure."[20] This also touched upon the code of **"against constitutional values"** which was only mentioned by 12 countries (33.33%).

Several terrorism definitions also included the three codes specifying the objectives of terrorism. 14 countries (38.89%) included **"support for political causes,"** 13 countries (36.11%) included **"support for ideological causes"** and 10 countries (27.78%) included **"support for religious causes."** These codes targeted whether or not countries specifically describe the motivations for terrorist acts and/or terrorism groups. Nine countries mentioned all of these codes, including Australia, Canada, Croatia, the Czech Republic, Kyrgyzstan, Pakistan, Trinidad and Tobago, the U.K., and Uzbekistan.

Violent Extremism

While some countries explicitly have a strategy and plan on countering terrorism, others either have a separate document on the strategy to counter violent extremism and/or relevant laws. This section provides an analysis of how the term "violent extremism" was defined in these NAPs and other documents. To recount from the methodology section, the 16 codes used in the above terrorism analysis are used in the analysis of the term "violent extremism."

[19] Czech Republic, Ministry of Justice of the Czech Republic, Criminal Code of Czech Republic, 2009, link, 136.
[20] Philippines, Government of the Philippines, National Action Plan on Preventing and CVE, 2019, document provided to Hedayah, 5.

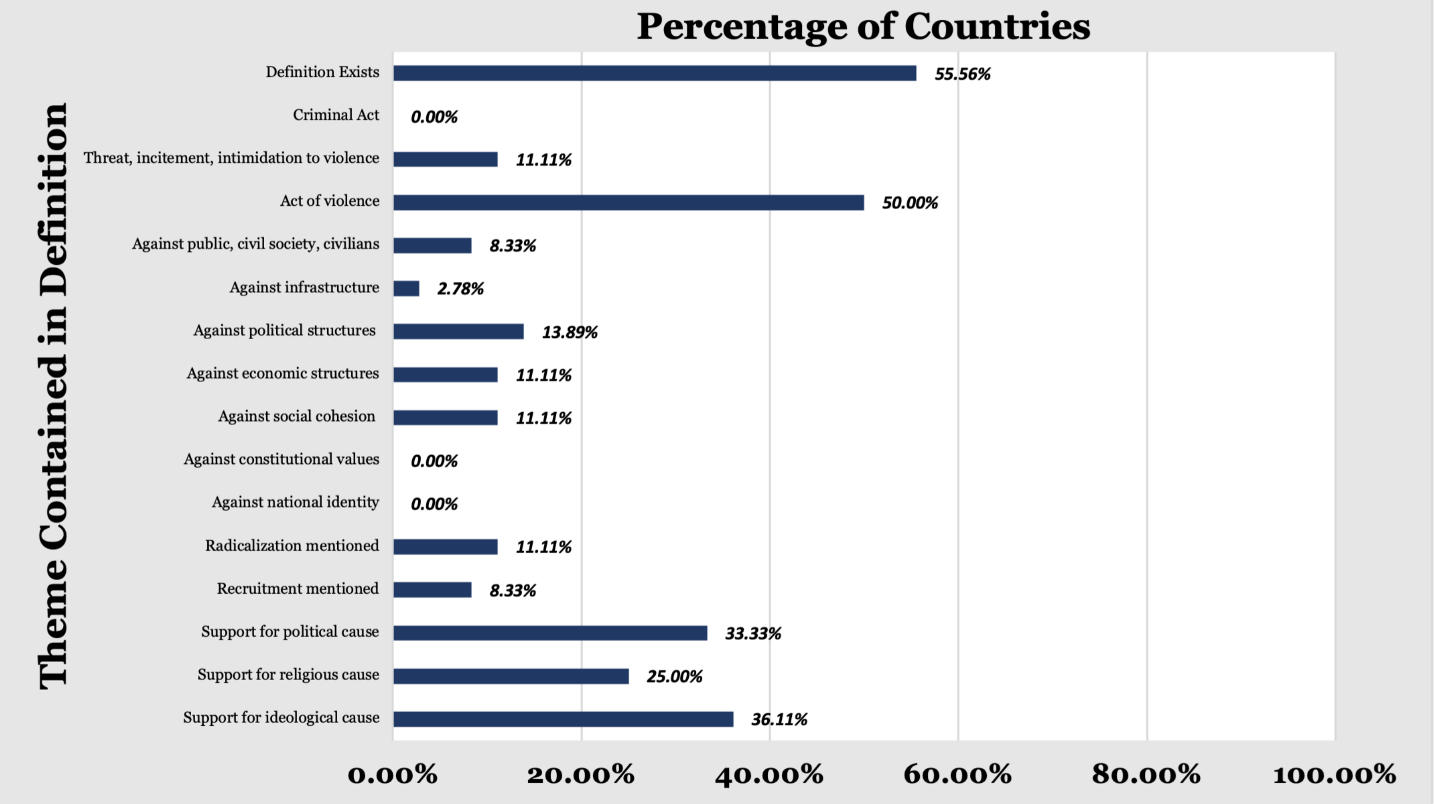


Figure 2. Definitions of Violent Extremism.

An assessment of relevant documents to identify the definition of violent extremism for each of the selected countries indicated that 20 countries (55.56%) had an existing definition of the term. "In some cases, countries did not explicitly define violent extremism and instead had a definition for "ideology of violence." Such was the case for Russia and Kazakhstan, with both of the countries' definitions for the term highlighting "Views and ideas that justify the use of violence to achieve political, ideological, [and] religious" goals.[21] As this closely reflected the definition of violent extremism in other countries, the term was equally considered.

As demonstrated in Figure 2, 18 out of the 20 countries with a definition (and 50% of the total number of countries analyzed) indicated the term related to **"acts of violence."** 13 countries explicitly mentioned **"support for ideological causes,"** 12 countries mentioned **"support for political causes"** and nine countries mentioned **"support for religious causes."** While the nature and the causes of violent extremism were mostly defined, few explicitly mentioned the target of these acts. Out of the 20 definitions, only five countries mentioned **"political structures"** as a target, four countries mentioned both "**economic structures,"** four countries mentioned **"social cohesion,"** three countries identified the **"public, civil society and civilians"** inclusive of the general population as a target, and one country included attacks **"against infrastructure."** Interestingly, no country had any

[21] Russia, Russian Federation, Strategy of Russian Federation for Countering Extremism until 2025, 2020, link. While Kazakhstan's definition does not follow the exact same wording, Russia's definition covers all those points also mentioned in Kazakhstan's.

mention of acts committed **"against constitutional values"** or **"against national identity."** Three countries (Canada, the Maldives and the Philippines) mention both **"radicalization"** and **"recruitment"** in their definition, with one other country (Belgium) only mentioning radicalization. Finally, the results showed that none of the countries' definitions of the term violent extremism had a mention of it being a **"criminal act."**

In looking at violent extremism it was not surprising to find the definitions varied across countries, and there were not any clear regional patterns or trends. One definition worth highlighting was the Philippines, which encompassed most of the aspects listed in the coding above. According to the Philippines NAP P/CVE document, violent extremism is:

> A belief system that drives individuals or groups to commit violent acts. This belief stems out of a context of repression, poverty and other so-called "push factors" made attractive by "pull factors" such as money, power, sense of purpose desired by the recruits, and charismatic VE leaders. The aim of VE is the furtherance of causes that are ideological, religious, political, social and/or economic in nature. It fosters hatred that may lead to intercommunity violence.[22]

[22] Philippines, National Action Plan, 5.

Extremism

The next term included in this research assessment is **"extremism."** Similar to violent extremism, some NAPs and relevant documents choose to refer to extremism as an umbrella term to include both violent and non-violent aspects, while others differentiate between the two. The definitions analyzed followed the same process as for violent extremism, with an additional analysis provided comparing the two. Analysis of the definitions of extremism indicated that only 16 countries (44.44%) had one in the relevant documents. In the case of Belgium, the term "radicalism" was considered in place of extremism as it closely covered similar points.

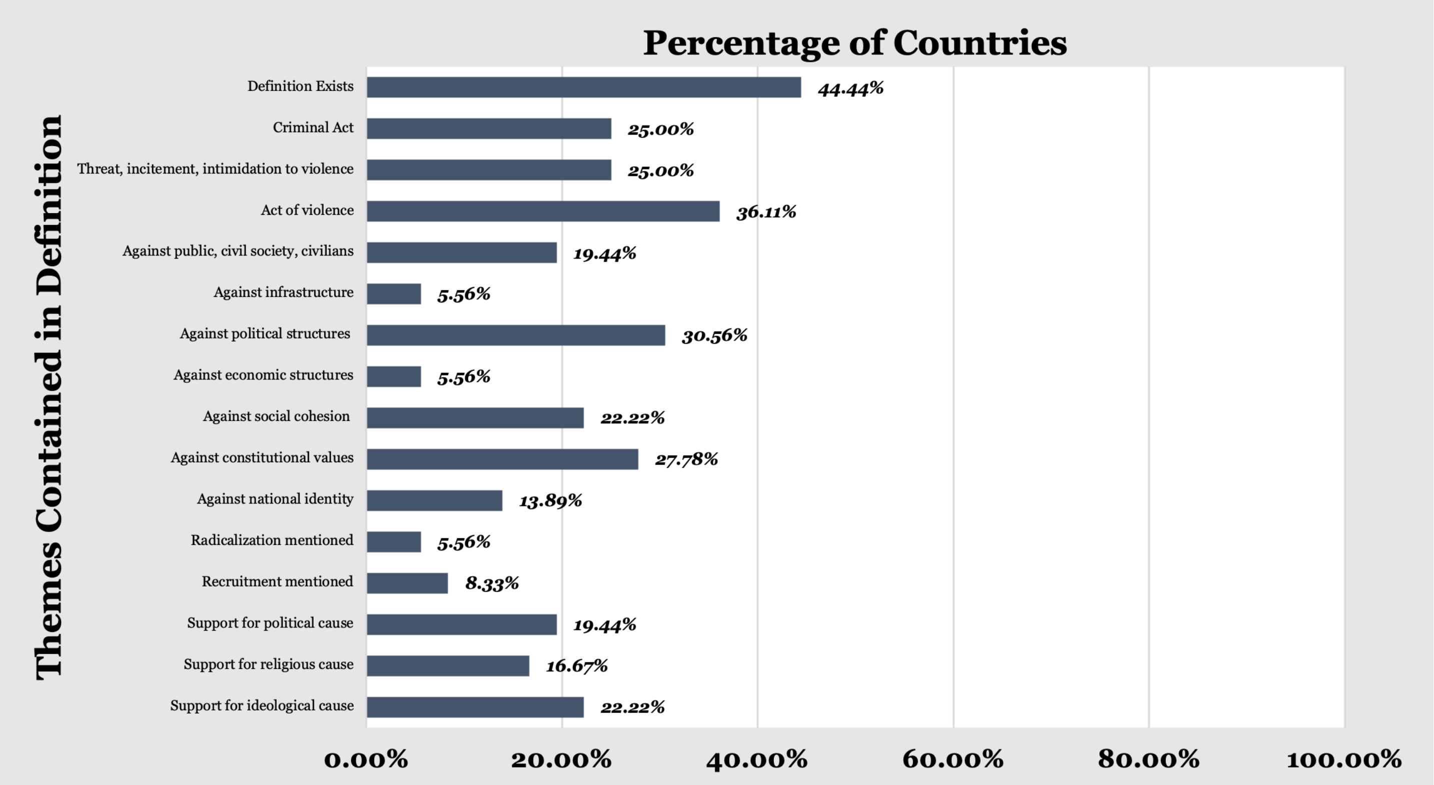


Figure 3. Definitions of Extremism.

As demonstrated in Figure 3, 13 countries out of the 16 which defined extremism (36.11% of the total number of countries) included in their definition an element of **"acts of violence"** (more specifically just "violence"). Of those 13, nine countries highlighted the **"threat, incitement, or intimidation to violence."** In assessing the mention of target audiences, the results indicate that all seven elements of the codes identified for this research were mentioned in various definitions: 11 countries mentioned the targeting of **"political structures,"** 10 countries mentioned attacks **"against constitutional values,"** eight countries mentioned **"social cohesion,"** seven countries reference attacks **"against public, civil society, or civilians,"** five countries mentioned **"against national identity,"** and two countries mentioned the targeting of **"infrastructure"** and **"economic structures."** Regarding the mention of **"radicalization" and "recruitment,"** two entities (the E.U. and Netherlands) had the explicit mention of the former and three entities (the E.U., Netherlands, and Tajikistan) included the latter. Finally, nine countries had a clear indication of extremism to be a **"criminal act."**

It is worth highlighting that although not a significant difference, in the definitions identified political structures were mentioned more often as the target of extremism than civilians as compared to the results seen in the coding of terrorism – where attacks against civilian populations were more frequently mentioned as targets than political structures. Furthermore, only five out of the nine definitions for extremism that mentioned **"support for political cause"** were correlated with those that mentioned targeting political structures.

Most of the countries did feature a comprehensive and extensive explanation of what the term extremism entails. For instance, the Czech Republic defined extremism as "Distinct ideological attitudes that deviate from constitutional, legal norms, are characterized by elements of intolerance, and attack basic democratic constitutional principles, as defined in the Czech constitutional order."[23] The Czech Republic's definition of extremism also identified some principles such as human rights, democratic state and principles, protection of minorities, and freedom and equality of people as aspects that extremism went against.

Comparison between violent extremism and extremism

It is noteworthy that out of the 36 countries, only five (13.88%) had a definition for both terms – Belgium, Bulgaria, Kazakhstan, Kosovo, and Russia. Furthermore, only two out of those five countries (Belgium and Kosovo) differentiated between one as "violent" and the other as simply "extreme," while the other 3 countries mentioned "violence" in both definitions.

It should also be noted that more countries had a definition of violent extremism in their documents than extremism. However, in a brief qualitative assessment of the definitions of the themes covered, the definitions of extremism were of higher quality and included more diverse themes than the definitions of violent extremism. While this report does not go into detail on a comprehensive qualitative assessment of these terms, it is recommended that countries incorporate diverse aspects of the terms into their definitions wherever possible to ensure clarity in how their country approaches the pheonoma.

Radicalization

As demonstrated in Figure 4, radicalization is defined in 17 countries (47.22%) from our case selection. The predominant code out of the seven codes[24] that we explored for this definition was "process." Interestingly, 16 countries referred to radicalization as a "process" – mainly explained as an individual or group experience. Although an overwhelming majority of these countries used this element to describe radicalization, there is variation in what the term "process" particularly entails. For instance, while some countries define it as a progression toward adopting radical religious beliefs or support for violent extremism,

[23] Czech Republic, Ministry of Interior of the Czech Republic, What is Extremism, 2010, link

[24] Codes for this section included "process," "networks," "extreme or radical opinions," "extreme or radical actions," "violence," "online," and "opportunity to intervene."

others perceive it as a cognitive or behavioral adaptation to an ideological worldview aiming at bringing about changes in society, or even a development that gives rise to threats against national security.

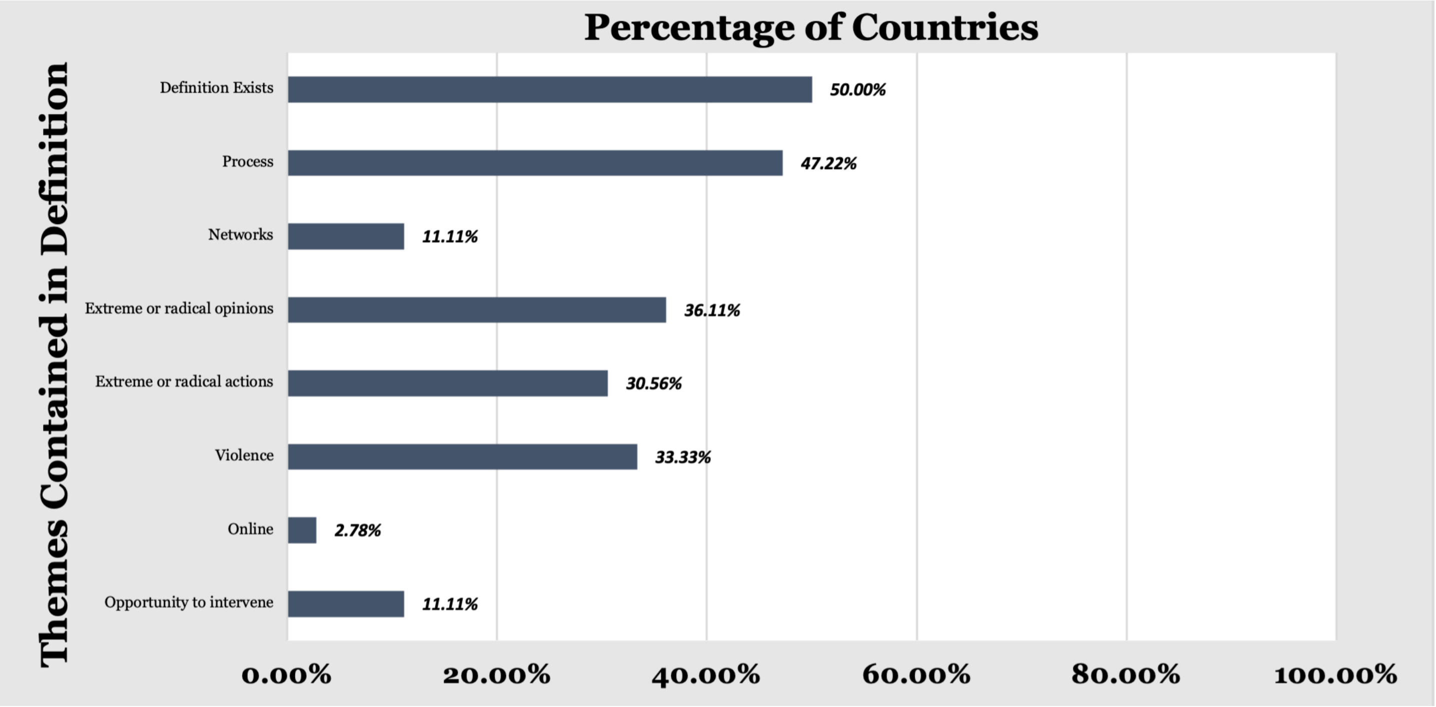


Figure 4. Definitions of Radicalization.

The second most common code that we analyzed was the reference to "extreme or radical opinions." This code was included in the definitions of 13 countries.The code often appeared in forms such as the imposing of opinions, subscribing to extremist views, or approving extremist beliefs. For instance, Belgium's NAP document described it as "[a] growing intolerance of the ideas of others."[25] The element of extreme or radical opinions is often mentioned as being against the mainstream views. The definition found in Canada's National Strategy on Countering Radicalization to Violence exemplifies this by suggesting that radicalization is when someone "[g]radually adopts extreme positions or ideologies that are opposed to the status quo and challenge mainstream."[26] Interestingly, for 12 out of the 13 definitions where this code was present, the code "process" was also present, implying that there is a significant correlation between the radicalization process and embracing extreme or radical opinions. While this code is relatively common in the definitions of radicalization, what is seemingly missing in the definitions is a more lucid explanation of the triggers and nature of one's progression from mainstream to radical views.

Furthermore, 11 countries correlated radicalization with **"extreme or radical actions."** In the definitions, countries often refer to extreme or radical actions in terms of one's support

[25] Belgium, Regierung der Deutschsprachigen Gemeinschaft Belgiens, Stratégie de Prévention du Radicalisme Violent en Communauté Germanophone de Belgique, 2019, link, 5.
[26] Canada, Government of Canada, National Strategy on Countering Radicalization to Violence, 2018, link, 9.

or acceptance of them (but not necessarily acting). For instance, Australia mentions in its documents that radicalization "Does not necessarily mean a willingness to use violence to realize those beliefs, but some individuals come to believe that violence is justified."[27]

The mention of **"violence"** was present in the definitions of 12 countries, mainly in the context of the support or justification of violence (though not necessarily being violent). Therefore, the important distinction made by countries between support for violence through radicalization and acting violently is apparent across the case selection.

The two codes that were less frequent in the data and were both found in only four countries were **"networks"** and **"opportunity to intervene."** The reference to "networks" was twofold. The term in the definitions was either used in the context of the need for supportive social networks to prevent radicalization or the detrimental effect social networks can have in radicalization and recruitment. In general, the neglect of this code in the data demonstrates that there is a potential need in the radicalization definitions to recognize the importance of social networks (especially online) in feeding into individual radicalization processes and as an important part of deradicalization efforts. On the other hand, the "opportunity to intervene" code included mentions of radicalization as reversible and the importance of the timing of prevention. For example, Bulgaria's Strategy to Counteract Radicalization and Terrorism document explicitly states that "In the earlier stages, radicalization is reversible and preventable."[28] This recognition of different stages of radicalization in the definition shows one layer of the complexity of radicalization as a phenomenon and the need to better understand it. When defining radicalization such perspectives must be considered.

Perhaps the most surprising finding for this definition was the near absence of the word **"online,"** which was the last code we tracked – especially when considering the general inclusiveness of the code. Although many legal documents extensively discuss radicalization in the context of online recruitment, only the definition of the European Union included this code. This was particularly interesting as we also expected it to arise in conjunction with other codes, especially the "process" code, resting on the premise that radicalization processes primarily occur online.

In sum, while there is an alignment of the codes and common themes arising in the definitions studied in our research, there is also an evident variation of elements constituting definitions of radicalization.

Preventing and Countering Violent Extremism

As Figure 5 demonstrates, a term related to CVE has been defined by 33 countries (91.67%

[27] Australia, Australian Government, Attorney-General's Department, Preventing Violent Extremism and Radicalisation in Australia, 2015, link, 28.

[28] Bulgaria, Republic of Bulgaria: Council of Ministers, Strategy to Counteract Radicalization and Terrorism (2015-2020), 2015, link, 2.

in our study. We explored three codes[29] that appear to be central in the definition of CVE (the definitions of the E.U. and five countries including Austria, Canada, Denmark, Indonesia, and Trinidad and Tobago contained all of these codes). Notably, the term for CVE was loosely defined—in some circumstances the country referred to "prevention" or "PVE," while in other circumstances the country may refer to "counter-radicalization." Regardless, defining CVE encompassed all related definitions for how the 36 entities in this study approached the prevention or countering of violent extremism and terrorism.

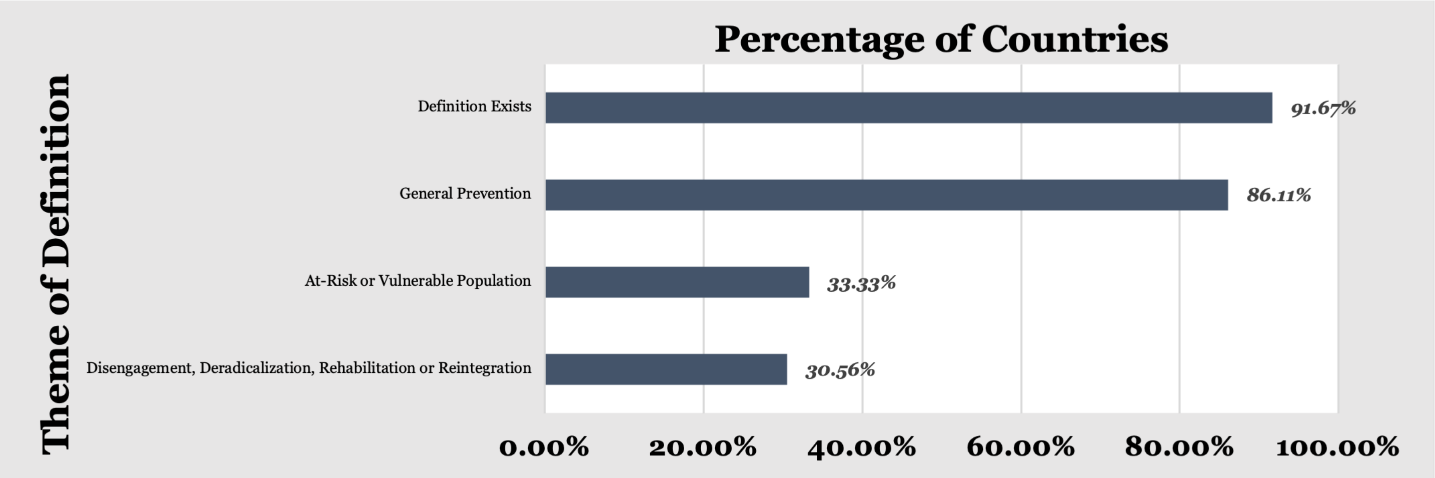


Figure 5. Definitions of Preventing and Countering Violent Extremism.

First, given that CVE is sometimes discussed in broad terms, the "general prevention" code was used to mean interventions that were aimed at a general population or building resilience to violent extremism.[30] 31 countries included general prevention initiatives in their definition. Some countries like Bosnia and Herzegovina focused on cooperation and collaboration between different sectors to prevent "[t]he processes of indoctrination in terrorist ideologies,"[31] while others such as the Albanian National Approach for CVE places a stronger emphasis on "[s]ocial, political, legal, education and economic programs" in their definition.[32] We also deemed it important to include a code that described the definitions that specifically used the terms **"at-risk," "vulnerable,"** or made specific reference to why a particular group was considered to be classified as such (and also references targeted detection). Definitions that included this code were found in the legal documents of 12 countries . For instance, the E.U.'s strategy clearly stated that P/CVE approaches should aim "To challenge radical or extremist communications at the platforms used most frequently by those who are most at risk to be radicalised."[33] Lastly, we included a code that encompassed **"disengagement, deradicalization, rehabilitation, and reintegration."**

[29] Keywords searched for this section included "general prevention," "at-risk" or "vulnerable" populations, and "disengagement, deradicalization, rehabilitation and reintegration."

[30] For more on how Hedayah characterizes different stages of P/CVE, see Cristina Mattei, The CVE Cycle: An Individual Trajectory, 2019, link.

[31] Bosnia and Herzegovina, Government of Bosnia and Herzegovina: Council of Ministers, Strategy of Bosnia and Herzegovina for Preventing and Combating Terrorism, 2015, link, 9.

[32] Albania, Albanian Council of Ministers, Albanian National Strategy for CVE, 2015, link, 5.

[33] European Union, Council of the European Union, Revised EU Strategy for Combating Radicalisation and Recruitment to Terrorism, 2014, link, 7.

In this code, we included elements from the definitions that included discussions of foreign terrorist fighters (FTFs), prisoners, and former violent extremists. This code was found in 11 countries. It is worth highlighting Trinidad and Tobago's definition that includes all these aspects:

> Preventing and CVE (PCVE) is essentially raising awareness and building the capacities and competencies of government and communities to become partners in collective efforts to prevent (stop people from becoming violent extremists or supporting terrorism in any form) and counter (for example, challenge extremist narratives and support sustainable disengagement and deradicalization efforts). Violent extremists often exploit social and economic conditions and individual vulnerabilities to recruit and motivate others. Efforts to prevent violent extremism therefore must address factors that make people vulnerable to extremist influences, including recruitment by terrorists.[34]

Internet and Social Media

The last important component that this research examined throughout the NAP documents was the emphasis on the internet and social media concerning radicalization, violent extremism, and appropriate responses. For this section, the national documents were searched for a list of keywords[35] and the relevant passages were noted for coding purposes.

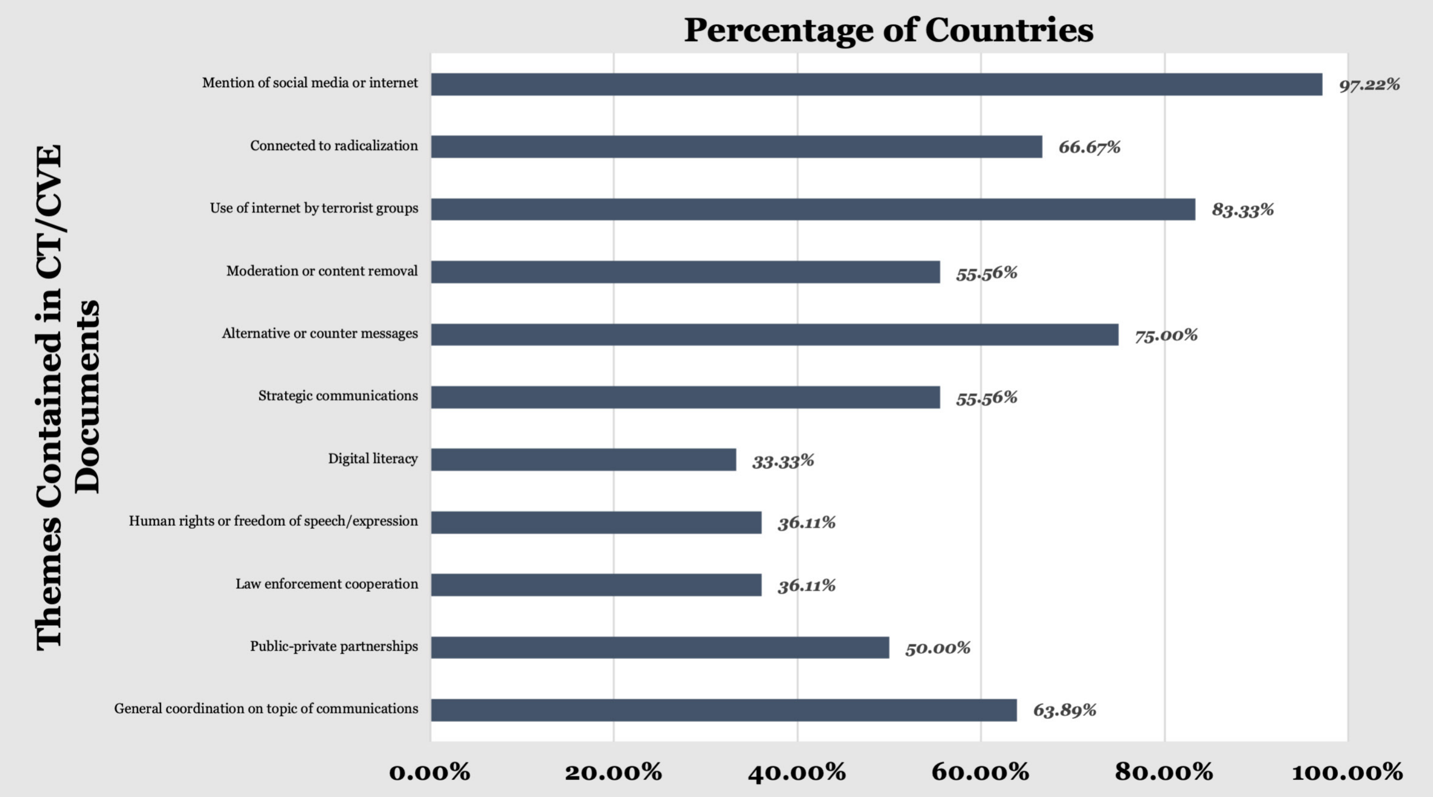


Figure 6. Representation of Internet and Social Media.

[34] Trinidad and Tobago, Ministry of National Security, Partnering to Safeguard Communities in Trinidad and Tobago National Policy and Guidelines for Preventing and CVE, 2019, 55.

[35] Keywords searched for this section included "social media," "internet," "private sector," "online," "counter-narrative/counter narrative," "counter-message/counter message," "propaganda," and "communications."

As the results in Figure 6 show, 2/3 of the countries (24 countries, 66.67%) recognize that radicalization occurs **"online"** and 30 countries (83.33%) mention that the **"internet"** is frequently used by terrorist groups to either recruit individuals or spread their **"propaganda."** For example, Canada states that "Violent extremist and terrorist organizations use the internet and social media in various ways… These include indoctrinating individuals into their ideologies and recruiting members to join their organizations or provide financial support. The online space is also used to inspire, incite, coordinate, finance and plan acts of violence."[36] However, it should be noted that in accordance with earlier discussions, the actual definitions of radicalization do not include the "online" component explicitly, and it is only through later contextual references that the online component of radicalization emerges. Still, these strategic documents recognize that the internet and social media are important factors in many P/CVE documents and critical to effectively countering the use of the internet by terrorist groups.

The countries studied in this research seem to place the heaviest weight on responses to the challenge of online radicalization that include **"alternative and counter messages"** – an element integrated into the NAP approaches of 27 countries (75%). For instance, one of Albania's strategic objectives is to "Reduce the impact of violent extremist propaganda and recruitment online by using social media to develop and disseminate alternative positive messages."[37] Austria notes that "Digital media can even represent an opportunity: They can be used effectively for campaigns organized in the fields of prevention of violent extremism and de-radicalization, for counselling victims and their reference persons as well as for the method of an alternative narrative."[38]

One of the ways in which **"counter narratives"** are often referenced in the documents in this study is with respect to key influencers, including youth, women, religious leaders, and social media influencers. For instance, Indonesia lists one strategic objective as the need to "Increase the participation of key youth, traditional leaders, religious figures and women in the media, social media companies, and social media influencers deliver the message to prevent radicalization that leads to terrorism."[39] Similarly, Nigeria states that "Social media influencers are critical to our counter messaging online. As violent extremists use the internet for recruitment and spreading of extremist ideas, we will continue to build networks of youths, students, community leaders who engage online to counter violent extremism."[40]

[36] Canada, National Strategy, 24.

[37] Albania, Albanian National Strategy, 7.

[38] Austria, Bundesweites Netzwerk Extremismus-pravention und Deradikalisierung, The Austrian Strategy for the Prevention and Countering of Violent Extremism and De-Radicalisation, 2018, link, 48.

[39] Indonesia, Office of the President of the Republic of Indonesia, Peraturan Presiden (PERPRES) tentang Rencana Aksi Nasional Pencegahan dan Penanggulangan Ekstremisme Berbasis Kekerasan yang Mengarah pada Terorisme Tahun 2020-2024 [Presidential Regulation Concerning the National Action Plan for Preventing and CVE that Leads to Terrorism in 2020-2024], 2021, link, 34.

[40] Nigeria, Federal Republic of Nigeria & The Office of the National Security Adviser, Policy Framework and National Action Plan for Preventing and CVE, 2017, link, 27.

Notably, 20 countries (55.56%) in this assessment explicitly mention the need to **"moderate or remove content"** from the online space. As an example, the Australian strategy mentions that "[i]nitiatives to counter violent extremism in online communication include working with social media companies on take-downs of violent extremist material, including through the Report Online Extremism tool for the public and increasing community capacity to undermine the appeal of violent extremism."[41]

"Strategic communications" — to inform the public or raise awareness about violent extremism, terrorism, or CVE — had an equal emphasis on content takedown among the countries in this research study, and 20 (55.56%) also mentioned this theme in their strategic documents. For example, the Czech Republic noted that "The public has the right to obtain from the relevant state institutions continuous information on the extremist scene and on anti-extremist activities. Informing about the extremist situation is also one of the tools for combating extremist entities."[42]
"Digital literacy" and enhancing young people's ability to think critically about information was also a potential response mentioned in the strategic documents, but only by 12 of the countries studied (33.33%). As an example, the Danish NAP states that:

> Educational material on how to be critical of sources, propaganda techniques and digital welfare and an online education package about using the internet and social media [are] under preparation… the objective is to sharpen children and young people's critical faculties, understanding of the digital media and their ability to see through and resist propaganda and extremist messages that they may come across on the internet and social media.[43]

The Maldives also mentions digital literacy in one of their main goals of "[f]ostering resilience to PCVE through awareness, counter narrative, digital literacy and critical thinking."[44]

Slightly worrisome was the relative paucity of mentions of respect for **"human rights and protecting freedom of speech"** or freedom of expression in the online space. Only 13 of the countries in this study (36.11%) included it as a part of their documents. Interestingly, this is also not portrayed consistently across all countries that mention this subject. For instance, Lebanon is one of the countries that places some responsibility on the technology sector, noting that the Ministry of Telecommunications should "monitor and prevent

[41] Australia, Council of Australian Governments, Australia's Counter-Terrorism Strategy: Strengthening Our Resilience, 2015, link, 19.
[42] Czech Republic, Ministry of Interior of the Czech Republic, Progress Evaluation of the Concept of Fighting Against Expressions of Extremism and Hate Prejudice for 2020, 2020, link, 3.
[43] Denmark, The Danish Ministry of Immigration, Integration and Housing, Preventing and Countering Extremism and Radicalisation: National Action Plan, 2016, link, 23.
[44] Maldives, National Counter Terrorism Centre & Government of the Maldives, National Action Plan for Preventing and CVE 2020-2024, 2020, link, 4.

extremist content on the internet by strategically collaborating with service providers and telecommunication companies to ensure that they are committed to preventing extremist accounts on the internet, deleting extremist content and modifying service delivery rules and contracts to ensure that participants adhere to human rights rules and refrain from posting any extremist content."[45] The Philippines' approachplaces a strong emphasis on the government's responsibility to respect human rights and privacy, stating, "To address problems of this nature, these [government] practitioners must have a thorough understanding of criminal, privacy and human rights law; data protection policies; and mutual legal assistance channels. Having knowledge and access to the up-to-date law enforcement guidelines of private communications service providers (CSPs) is also essential."[46]

Finally, there are several ways in which the strategic documents identify areas of collaboration on the challenges related to radicalization and the spread of terrorist content on the internet and social media. 13 countries (36.11%) mention the need for **"law enforcement to cooperate with the private sector"** or the technology industry on the subject of removing terrorist content. For example, according to U.K. documents, "We are already working in partnership with industry and the police to remove terrorist and extremist material. Cooperation with industry has significantly improved in recent years."[47] 18 countries (50%) mention a need to cooperate between **"public entities (governments) and the private sector**" when it comes to communications activities. The Canadian strategy states that "It is critical for governments to cooperate with technology companies to effectively address violent extremists and terrorist use of the internet."[48] (Notably, the same strategy also explicitly mentions GIFCT as a critical partner in this endeavor.) Finally, 23 countries (63.89%) state some sort of **"general cooperation"** on the subject of social media and the internet, even if the technology sector or private sector are not explicitly mentioned. For example, with reference to countering Al-Shabaab ideologies, Somalia states that "International good practice demonstrates that often the institutions of government alone are not necessarily the most effective communicators in this regard. It is therefore critical that credible partners are identified who can serve as influencers to those who are most vulnerable to being drawn toward violent extremism."[49] There is a recognition that there are other critical stakeholders needed to prevent the spread of Al-Shabaab ideologies, including online, other than government entities.

Recommendations

In terms of better defining and classifying terrorist content in the online space, these results

[45] Lebanon, Lebanese Republic: Presidency of the Council of Ministers, National Strategy for Preventing Violent Extremism, 2018, link, 65.
[46] Philippines, National Action Plan, 58.
[47] United Kingdom, Her Majesty's Government, Counter-Extremism Strategy, 2015, link, 24.
[48] Canada, National Strategy, 27-28.
[49] Somalia, Republic of Somalia, National Strategy and Action Plan for Preventing and CVE, 2016, document provided to Hedayah.

reveal several challenges and opportunities for the technology sector hoping to define their activities around preventing terrorist content on their platforms. The recommendations for GIFCT members and the broader technology sector stemming from these results are as follows:

1. **Digital platforms and the technology sector should seek to define terms for themselves as they apply to their unique platforms globally.** Third-party definitions of terrorism, violent extremism, extremism and radicalization coming from governments are not consistent or reliable even within one region. While some countries have more robust definitions of these terms, other countries are vague in their approaches, especially when it comes to terrorist and violent extremist activities online. This means that the technology sector will need to lead the way in defining terrorist content online because current government approaches are not comprehensive enough to define the practical implications of how terrorism manifests online. In this regard, each organization should assess the existing definitions separately and formulate ones that can be implemented in their online sphere.
2. **Digital platforms that work internationally should leverage the best aspects of the definitions of these terms to find a global definition that is broad enough to be applied nationally.** This study also found that there are few regional trends and nuances to the definitions of these terms, which means that it may be best to set a global definition of community standards that is adapted to national frameworks and contexts as opposed to regional frameworks.
3. **Emphasize the importance of the "threat, incitement and intimidation" factor in terrorism definitions.** Because these aspects are often related to social media and the internet, it is extremely important that digital platforms define how (online) threats/incitements relate to committing acts of terrorism.
4. **Align interests of digital platforms with national interests.** The most frequently used and cited codes mentioned in this study are the aspects of definitions that national governments care about the most. Those elements should be incorporated into the community standards and guidelines for the technology sector. In doing so, digital platforms will be able to use their definitions as leverage for cooperation with governments moving forward on this topic.
5. **Advocate to national governments the inclusion of specific roles and responsibilities for digital platforms in their NAPs,** and to define what that relationship means with respect to content removal, counter narratives, strategic communications, and other responses. Although cooperation in a general sense on the subject of online radicalization and recruitment is mentioned in these strategic documents by almost 2/3 of the countries in this study, only half explicitly identify the technology sector as critical stakeholders. This means that there is also an opportunity for the technology sector to lobby for the need for increased cooperation between governments and the technology sector. Since many countries are currently or will soon review and revise their national strategies, urgent engagement and cooperation on this topic

is needed.

6. **Ensure individual companies are clear about their own company values towards human rights and privacy,** and that their actions in a country both abide by local law and their own values and ethics. The technology sector as a whole is concerned with ethical implications of their definitions and activities related to terrorist content removal, specifically with respect to human rights, freedom of expression, freedom of speech, and open internet. However, as the results show, the emphasis on human rights is not as strong in the national strategic documents related to countering terrorism and violent extremism, so this is an area in which the technology sector can play a significant role in consulting with governments to define the parameters of what is classified as terrorist content and what is not. Importantly, the technology sector should work with governments to explain specific ethical situations and decisions that they may face (while protecting private data), and maintain good communication on where the line can and should be drawn with respect to content being available online. Of course, one of the main ethical challenges to labeling certain content as "terrorist" content is the risk that political opposition parties and their narratives are limited or censored (intentionally or unintentionally).
7. **Share information about practical challenges to defining terrorism in the online space.** Based on the definitions included in this study, governments in general recognize the importance of the internet in terms of radicalization and recruitment to terrorism (as well as prevention). However, there is a limited understanding of the challenges that technology companies face in this field. Forming working groups and strategic partnerships that include informing policymakers of the practical challenges and solutions would aid in filling this gap between policy and implementation.

Annex A: List of Countries and Documents

Click here for a list of countries and documents.

Bibliography

Albania (2015), Albanian National Strategy for Countering Violent Extremism (Albania: Albanian Council of Ministers), link

Australia (2015), Australia's Counter-Terrorism Strategy: Strengthening Our Resilience (Australia: Council of Australian Governments), link

Australia (2015), Preventing Violent Extremism and Radicalisation in Australia (Australia: Australian Government, Attorney-General's Department),

Austria (2018), The Austrian Strategy for the Prevention and Countering of Violent Extremism and De-Radicalisation (Austria: Bundesweites Netzwerk Extremismus-pravention und Deradikalisierung), link

Belgium (2019), Stratégie de Prévention du Radicalisme Violent en Communauté Germanophone de Belgique (Belgium: Regierung der Deutschsprachigen Gemeinschaft Belgiens), link

Bosnia and Herzegovina (2015), Strategy of Bosnia and Herzegovina for Preventing and Combating Terrorism (Bosnia and Herzegovina: Government of Bosnia and Herzegovina: Council of Ministers), link

Bulgaria (2015), Strategy to Counteract Radicalization and Terrorism (2015-2020) (Bulgaria: Republic of Bulgaria: Council of Ministers), link

Canada (2018), National Strategy on Countering Radicalization to Violence (Canada: Government of Canada), link

Croatia (2015), National Strategy for Prevention and Suppression of Terrorism (Croatia: Government of the Republic of Croatia), link

Czech Republic (2009), Criminal Code of Czech Republic (Czech Republic: Ministry of Justice of the Czech Republic), link

Czech Republic (2020), Progress evaluation of the concept of fighting against expressions of extremism and hate prejudice for 2020 (Czech Republic: Ministry of Interior of the Czech Republic), link

Cezch Republic (2010), What is Extremism? (Czech Republic: Ministry of Interior of the Czech Republic), link

Denmark (2016), Preventing and Countering Extremism and Radicalisation: National Action Plan (Denmark: The Danish Ministry of Immigration, Integration and Housing), link

European Union (2014), Revised EU Strategy for Combating Radicalisation and Recruitment to Terrorism (European Union: Council of the European Union) link

Indonesia (2021), Peraturan Presiden (PERPRES) tentang Rencana Aksi Nasional Pencegahan dan Penanggulangan Ekstremisme Berbasis Kekerasan yang Mengarah pada Terorisme Tahun 2020-2024 [Presidential Regulation Concerning the National Action Plan for Preventing and CVE that Leads to Terrorism in 2020-2024] (Indonesia: Office of the President of the Republic of Indonesia), link

Lebanon (2018), National Strategy for Preventing Violent Extremism (Lebanon: Presidency of the Council of Ministers), link

Maldives (2020), National Action Plan for Preventing and CVE 2020-2024 (Maldives: National Counter Terrorism Centre & Government of the Maldives), link

Mattei, Cristina (2019), The CVE Cycle: An Individual Trajectory (Abu Dhabi: Hedayah), link

Netherlands (2012), Criminal Code of the Kingdom of Netherlands (Netherlands: Government of the Kingdom of Netherlands), link

Nigeria (2017), Policy Framework and National Action Plan for Preventing and CVE (Nigeria: Federal Republic of Nigeria & The Office of the National Security Adviser), link

Philippines (2019), National Action Plan on Preventing and Countering Violent Extremism (Philippines: Government of the Philippines), Document provided to Hedayah.

Russia (2020), Strategy of Russian Federation for Countering Extremism until 2025 (Russia: Russian Federation), link

Somalia (2016), National Strategy and Action Plan for Preventing and CVE (Somalia: Republic of Somalia), Document provided to Hedayah.

Trinidad and Tobago (2019), Partnering to Safeguard Communities in Trinidad and Tobago National Policy and Guidelines for Preventing and CVE (Trinidad and Tobago: Ministry of National Security), Document provided to Hedayah.

Turkmenistan (2014), Criminal Code of Turkemnistan (Turkemistan: Ministry of Adalat of Turkmenistan), link

United Kingdom (2015), Counter-Extremism Strategy (UK: Her Majesty's Government), link

United Nations (2015), Plan of Action to Prevent Violent Extremism (UN: Report to the Secretary General), link

Conclusion & Initial Next Steps

By GIFCT Staff

Conclusion & Initial Next Steps

As GIFCT develops a foundation for expanding its taxonomy for hash-sharing, GIFCT recognizes the need for a multi-stakeholder approach and appreciates the compelling research papers in this volume as well as the in-depth feedback from the GIFCT member companies, GIFCT's Operating Board, its Independent Advisory Committee, and others.

The research papers in this collection recognize that there are fundamentally two approaches to taxonomy expansion that can build on the existing system and complement its strengths and weaknesses. First, GIFCT could expand its taxonomy based on a behavioral and content-focused approach (as GIFCT does with hashes related to Content Incident Protocols). Second, GIFCT could expand its taxonomy based on organizations, through a list-based approach (as GIFCT does with the 93 percent of its hashes associated with U.N.-designated entities).

In addition, the recent GIFCT Human Rights Impact Assessment recommends that GIFCT "accompany the expansion of the hash-sharing database to include violent extremist content adequate transparency and oversight mechanisms."[50] Consistent with this recommendation and its own organizational values, GIFCT will continue to prioritize efforts to introduce greater transparency to GIFCT's work across all work streams, including the hash-sharing effort.

Feedback from GIFCT member companies clearly showed that GIFCT also has a pressing need to improve the breadth of capabilities that it provides to support its members in preventing terrorists and violent extremists from exploiting their platforms. The hash-sharing effort is a foundational and critically important component of the support that GIFCT provides to member companies to assist with their own efforts to surface, review, and remove terrorist and violent extremist content, but GIFCT's value to member companies must continue to grow and expand beyond the database. Any new approaches GIFCT undertakes must focus on developing solutions that both recognize how terrorist and violent extremist activities manifest online and are explainable, practical, and scalable.

Developed from the feedback of experts and interviews with tech companies, GIFCT will expand its taxonomy in a careful and deliberate manner, based on iterative steps rooted in increased transparency. In this way GIFCT can incrementally diversify what is available in the database, addressing the inherent bias towards inclusion of Islamist extremist terrorist content in its current list-based approach while carefully assessing its impact on human rights.

Based on the feedback GIFCT received and its assessment of what is feasible given the current architecture of the database, GIFCT intends as an initial step to expand the hash-

[50] Business for Social Responsibility, "Human Rights Assessment: Global Internet Forum to Counter Terrorism," (2021), link

sharing database to include the following categories of hashed content that reflect how terrorist and violent extremist content manifests online:

1. **Attacker Manifestos:** Hashed PDFs of violent extremist and terrorist attacker manifestos;
2. **Branded Terrorist Publications:** Hashed PDFs of branded terrorist publications; and
3. **TCAP URLs:** Hashed URLs corresponding with terrorist content links flagged to companies through Tech Against Terrorism's Terrorist Content Analytics Platform (TCAP).

Attacker Manifestos: There have been numerous cases of attackers that have posted their manifestos online in advance of carrying out attacks. Rarely are these individuals formally members of a known terrorist organization, but their manifestos are shared widely online by those praising, supporting, and inciting further hate-based violence. Researchers of white supremacy and neo-Nazi organizations have cited the proliferation of these documents and how they serve as a beacon to sympathizers of the attackers.[51] Hashed images and hashed text extracted from PDFs of violent extremist and terrorist attacker manifestos will be included within the database. This will allow GIFCT to expand beyond lists in a deliberate manner with well-defined parameters.

Branded Terrorist Publications: Branded content offers tech companies clear indicators that an image or video is in fact terrorist content. These publications are also developed with the specific aim of reaching wider audiences online – communicating with existing members and recruiting new members. To date, the hash-sharing database focuses on image and video hashes; however, most branded terrorist content is in PDF form. GIFCT will add the hashes of the text and images from these PDFs to the database, allowing GIFCT to expand its capabilities to deal with terrorist and violent material beyond images and videos.

TCAP URLs: At the end of 2020 Tech Against Terrorism launched TCAP. As part of its TCAP efforts, Tech Against Terrorism flags URLs relating to terrorist content to the tech company that hosts the infrastructure being used, similar to an Internet Referral Unit. They are clear about which designated terrorist groups are included in their efforts, incorporating ISIS and Al-Qaeda organizations and affiliates as well as Five Eyes designated far-right terrorist organizations.[52] As indicated by many GIFCT member companies, URLs are a key signal for companies. Terrorist content is often hosted on one platform and amplified on

[51] See A. Mattheis, "Manifesto memes: The radical right's new dangerous visual rhetorics – Centre for Analysis of the Radical Right," Centre for Analysis of the Radical Right, September 18, 2019, link; A. Romano, "The Christchurch shooter's manifesto used memes to spread hate," Vox (Blog), March 16, 2019, link; L. Dearden, "Revered as a saint by online extremists, how Christchurch shooter inspired copycat terrorists around the world," The Independent, August 24, 2019, link.

[52] Tech Against Terrorism, "Group Inclusion Policy," Terrorism Content Analytics Platform, retrieved May 10, 2021, link.

another.[53] In an effort to build on the utility and impact of TCAP efforts, GIFCT will include hashes of URLs that TCAP has flagged to tech companies. By hashing the URLs, GIFCT can ensure that no personally identifiable information is hosted or shared by GIFCT while allowing hashes of URLs to be shared as a signal to all GIFCT members.

There remain significant questions to be answered as GIFCT grows. Should GIFCT build its capacities to manage a list of verified violent extremist organizations that goes beyond existing government frameworks? How can GIFCT ensure wider transparency and quality control of hashes within the database, especially as newer companies join? How can GIFCT evolve other tools and approaches to cross-platform counter-terrorism and CVE efforts as threats evolve? How can GIFCT ensure that it effectively embeds human rights principles in these efforts?

Needless to say, the conversation is not over, nor will all these questions be answered without continuing to lean into GIFCT's global stakeholder community through its Independent Advisory Committee, working groups, and other channels. GIFCT hopes that these three focused expansions within the hash-sharing database taxonomy show meaningful evolution in line with strategies and recommendations put forward by the authors within this collection of research while taking a thoughtful and deliberate approach to terrorist and violent extremist exploitation of the Internet.

[53] This was well documented by the GIFCT-funded Global Research Network on Terrorism and Technology (GRNTT) and in particular in S. Macdonald, D. Grinnell, and N. Lorenzo-Dus, "A Study of Outlinks Contained in Tweets Mentioning Rumiyah," Global Research Network on Terrorism and Technology 2 (2019), link.

GIFCT
Global Internet Forum
to Counter Terrorism

Appendix:

Definitions of Terrorism & Violent Extremism

By Nayanka Paquete Perdigao, Sarah Kenny, and Aaron Tielemans

Appendix I:

Definitions of Terrorism & Violent Extremism

By Nayanka Paquete Perdigao, Sarah Kenny and Aaron Tielemans

Table of Contents

Introduction

As an independent NGO, GIFCT supports and develops research, guidance and tools to tech companies who are committed to cross-industry efforts to counter the spread of terrorist and violent extremist activity online. As part of this wider collection of briefing papers on the feasibility of expanding the taxonomy of the GIFCT hash-sharing database, this document serves as a resource and repository of select global definitions of terrorism and violent extremism. This includes relevant definitions to GIFCT members and its multi-stakeholder engagement and governance.

To date, GIFCT allows for a broad interpretation and discussion of terrorism and violent extremism within its programmatic and research efforts, but a narrow definition for inclusion in the hash-sharing database, since hashes have the potential to lead to source content on a given platform with possible repercussions for the user who shared or stored the content.

There is no universal agreement on the definition of terrorism, and far less agreement on the parameters and legal definitions of violent extremism. Tech companies developing policies on terrorism and violent extremism often ask which definitions to apply or which list-based approaches to employ. While most publicly accessible national and international lists have shortcomings, as noted by several authors in this collection of research, they are often the primary resource for smaller and newer companies tackling counter-terrorism and counter-extremism for the first time.

Despite the lack of a universally accepted definition of terrorism and violent extremism, this appendix brings together some of GIFCT's immediate stakeholders' approaches to understanding and responding to the exploitation of digital platforms by terrorists and violent extremists. We chose to provide definitions attributed to the government affiliated bodies on GIFCT's Independent Advisory Committee as well as by GIFCT's founding member companies.

Dictionary Definitions

Oxford University Press

Terrorism

NOUN: The unlawful use of violence and intimidation, especially against civilians, in the pursuit of political aims.[54]

Extremism

NOUN mass noun: The holding of extreme political or religious views; fanaticism.[55]

[54] Oxford University Press, s.v. "Terrorism," accessed April 21, 2021, link.

[55] Oxford University Press, s.v. "Extremism," accessed April 21, 2021, link.

Merriam-Webster

Terrorism

NOUN: the systematic use of terror especially as a means of coercion.[56]

Extremism

NOUN: The quality or state of being extreme; Advocacy of extreme measures or views.[57]

IAC Government & Regional Definitions

The government definitions in this section were chosen due to their role on the GIFCT Independent Advisory Committee (IAC). The IAC guides the Operating Board, including producing an annual report advising on organizational priorities and reflecting on previous performance. The IAC includes members from government and intergovernmental entities.[58]

Canada

According to the Canadian Criminal Code, "terrorist activity" means:

1. **an act or omission that is committed in or outside Canada and that, if committed in Canada, is one of the following offences:**

   b. the offences referred to in subsection 7(2) that implement the Convention for the Suppression of Unlawful Seizure of Aircraft, signed at The Hague on December 16, 1970,

   c. the offences referred to in subsection 7(2) that implement the Convention for the Suppression of Unlawful Acts against the Safety of Civil Aviation, signed at Montreal on September 23, 1971,

   d. the offences referred to in subsection 7(3) that implement the Convention on the Prevention and Punishment of Crimes against Internationally Protected Persons, including Diplomatic Agents, adopted by the General Assembly of the United Nations on December 14, 1973,

   e. the offences referred to in subsection 7(3.1) that implement the International Convention against the Taking of Hostages, adopted by the General Assembly of the United Nations on December 17, 1979,

   f. the offences referred to in subsection 7(2.21) that implement the Convention on the Physical Protection of Nuclear Material, done at Vienna and New York on March 3, 1980, as amended by the Amendment to the Convention on the

[56] Merriam-Webster, s.v. "Terrorism," accessed April 21, 2021, link.
[57] Merriam-Webster, s.v. "Extremism," accessed June 1, 2021, link.
[58] "Governance," GIFCT, 2021, link.

Physical Protection of Nuclear Material, done at Vienna on July 8, 2005 and the International Convention for the Suppression of Acts of Nuclear Terrorism, done at New York on September 14, 2005,

g. the offences referred to in subsection 7(2) that implement the Protocol for the Suppression of Unlawful Acts of Violence at Airports Serving International Civil Aviation, supplementary to the Convention for the Suppression of Unlawful Acts against the Safety of Civil Aviation, signed at Montreal on February 24, 1988,

h. the offences referred to in subsection 7(2.1) that implement the Convention for the Suppression of Unlawful Acts against the Safety of Maritime Navigation, done at Rome on March 10, 1988,

i. the offences referred to in subsection 7(2.1) or (2.2) that implement the Protocol for the Suppression of Unlawful Acts against the Safety of Fixed Platforms Located on the Continental Shelf, done at Rome on March 10, 1988,

j. the offences referred to in subsection 7(3.72) that implement the International Convention for the Suppression of Terrorist Bombings, adopted by the General Assembly of the United Nations on December 15, 1997, and

k. the offences referred to in subsection 7(3.73) that implement the International Convention for the Suppression of the Financing of Terrorism, adopted by the General Assembly of the United Nations on December 9, 1999, or

2. an act or omission, in or outside Canada,

a. that is committed

i. in whole or in part for a political, religious or ideological purpose, objective or cause, and

ii. in whole or in part with the intention of intimidating the public, or a segment of the public, with regard to its security, including its economic security, or compelling a person, a government or a domestic or an international organization to do or to refrain from doing any act, whether the public or the person, government or organization is inside or outside Canada, and

b. that intentionally

i. causes death or serious bodily harm to a person by the use of violence,

ii. endangers a person’s life,

iii. causes a serious risk to the health or safety of the public or any segment of the public,

iv. causes substantial property damage, whether to public or private property, if causing such damage is likely to result in the conduct or harm referred to in any of clauses (i) to (ii), or

v. causes serious interference with or serious disruption of an essential service, facility or system, whether public or private, other than as a result of advocacy,

> protest, dissent or stoppage of work that is not intended to result in the conduct or harm referred to in any of clauses (i) to (ii),and includes a conspiracy, attempt or threat to commit any such act or omission, or being an accessory after the fact or counselling in relation to any such act or omission, but, for greater certainty, does not include an act or omission that is committed during an armed conflict and that, at the time and in the place of its commission, is in accordance with customary international law or conventional international law applicable to the conflict, or the activities undertaken by military forces of a state in the exercise of their official duties, to the extent that those activities are governed by other rules of international law.[59]

Terrorist offences are also defined by the Anti-terrorism Act, 2015 (also known as Bill C-51) which is an act of the Parliament of Canada that broadened the authority of Canadian government agencies to share information about individuals easily.[60]

Violent Extremism

Within the broader designation of "violent extremism," Canada demarcates between religiously motivated, politically motivated, and ideologically motivated violent extremism as follows:

Religiously Motivated Violent Extremism (RMVE)

Ideologies that underpin RMVE often cast an individual as part of a spiritual struggle with an uncompromising structure of immorality. RMVE ideologies assure their adherents that success or salvation — either in a physical or spiritual realm can only be achieved through violence.

Politically Motivated Violent Extremism (PMVE)

PMVE narratives call for the use of violence to establish new political systems – or new structures and norms within existing systems. Adherents focus on elements of self-determination or representations rather than concepts of racial or ethnic supremacy.

Ideologically Motivated Violent Extremism (IMVE)

IMVE is often driven by a range of grievances and ideas from across the traditional ideological spectrum. The resulting worldview consists of a personalized narrative which centers on an extremist's willingness to incite, enable and or mobilize to violence. Extremists draw inspiration from a variety of sources including books, images, lectures, music, online discussions, videos and conversations.[61]

[59] "Justice Laws Website," Legislative Services Branch, 2019, link.
[60] "Understanding definitions of terrorism," Briefing European Parliamentary Research Service, 2015, link.
[61] "Threats to the security of Canada and Canadian Interests," Service, C. S. I., May 20, 2020, link.

France

According to the French Legislation, Article 421-1 Amended by Law No. 2005-1550 of December 12, 2005 – Art. 17, the following offences, when they are intentionally associated with an individual or collective undertaking committed with the express intention of gravely undermining public order by the use of intimidation or terror, constitute acts of terrorism:

1. deliberate attacks upon life;
2. deliberate attacks on integrity of the person; abduction, holding of persons against their will;
3. hijacking of an aircraft, ship or other means of transport; theft, extortion, destruction of and damage to property;
4. computer offences (as defined in Section III of the Criminal Code);
5. offences involving prohibited combat groups and movements;
6. offences involving firearms, explosives or nuclear substances;
7. handling the proceeds of one of the above offences;
8. money laundering;
9. insider offences;
10. endangering human, animal or environmental health by introducing substances into the air, soil, subsoil, foodstuffs or foodstuff ingredients, or water.[62]

Apart from criminal law, which is the main legal weapon against terrorism, French anti-terrorist law also draws on civil and administrative law (interception for security reasons, refusal of admission, refusal of asylum, refusal of admission, refusal of asylum, refusal of naturalization, deprivation of rights, expulsion, removal, supervision of associations, combat groups and private militias, and freezing of assets).

Ghana

According to the Ghana Anti-Terrorism Act, Section 2:

1. An act is a terrorist act if it is performed in furtherance of a political, ideological, religious, racial or ethnic cause and
   a. causes serious bodily harm to a person;
   b. causes serious damage to property;
   c. endangers a person's life;
   d. creates a serious risk to the health or safety of the public;

[62] "Code Pénal," France, French Government, Link.

   e. involves the use of firearms or explosives;

   f. releases into the environment or exposes the public to

      i. dangerous, hazardous, radioactive or harmful substances;

      ii. toxic chemicals; or

      iii. microbial or other biological agents or toxins;

   g. is prejudicial to national security or public safety;

   h. is designed or intended to disrupt

      i. a computer system or the provision of services directly related to communications;

      ii. banking or financial services;

      iii. utilities, transportation; or

      iv. other essential services; or

      v. is designed or intended to cause damage to essential infrastructure.

2. A person who contravenes subsection (1) commits an offence and is liable on conviction on indictment to a term of imprisonment of not less than seven years and not more than twenty-five years. [63]

Japan

The Ministry of Justice provides a definition for "Crime of Preparation of Acts of Terrorism and Other Organized Crimes" which focuses on the following elements:

> (1) an organized criminal group (2) planned serious crimes and (3) made preparations to carry out those plans. This makes it possible to arrest the criminals before any crime is actually carried out, which can prevent harm and other forms of damage.[64]

Examples of cases that constitute the Crime of Preparation of Acts of Terrorism and Other Organized Crimes include:

1. A terrorist group planned to manufacture deadly chemical substances, planned the mass-murder of civilians, and acquired a part of the necessary ingredients for such substances.
2. A terrorist group planned to hijack several planes and use them to attack skyscrapers, and for example, booked tickets for the planes.

[63] N. Yamoah, Anti -Terrorism Act, 2008 (ACT 762), Ghana Justice, August 14, 2020, link.

[64] "Crime of Preparation of Acts of Terrorism and Other Organized Crimes," Japan, Ministry of Justice, 2020, link.

3. A terrorist group planned to develop a computer virus and planned to utilize the computer virus to cause malfunctions in the electronic control systems of electric, gas, and water companies across the nation to paralyze large cities' vital infrastructure systems and cause panic, and started to develop such computer virus.[65]

New Zealand

According to the New Zealand's Terrorism Suppression Act 2002:

From Section 4, a designated terrorist entity means an entity:

1. for the time being designated under section 20 or 22 as a terrorist entity or associated entity; or
2. that is a United Nations listed terrorist entity

From Section 5,

1. an act is a terrorist act for the purposes of this Act if—
   a. the act falls within subsection (2); or
   b. the act is an act against a specified terrorism convention (as defined in section 4(1)); or
   c. the act is a terrorist act in armed conflict (as defined in section 4(1)).
2. An act falls within this subsection if it is intended to cause, in any 1 or more countries, 1 or more of the outcomes specified in subsection (3), and is carried out for the purpose of advancing an ideological, political, or religious cause, and with the following intention:
   a. to induce terror in a civilian population; or
   b. to unduly compel or to force a government or an international organization to do or abstain from doing any act.
3. The outcomes referred to in subsection (2) are—
   a. the death of, or other serious bodily injury to, 1 or more persons (other than a person carrying out the act):
   b. a serious risk to the health or safety of a population:
   c. destruction of, or serious damage to, property of great value or importance, or major economic loss, or major environmental damage, if likely to result in 1 or more outcomes specified in paragraphs (a), (b), and
   d. serious interference with, or serious disruption to, an infrastructure facility, if

[65] "Crime of Preparation," Japan, Ministry of Justice.

likely to endanger human life:

e. introduction or release of a disease-bearing organism, if likely to devastate the national economy of a country.

4. However, an act does not fall within subsection (2) if it occurs in a situation of armed conflict and is, at the time and in the place that it occurs, in accordance with rules of international law applicable to the conflict.
5. To avoid doubt, the fact that a person engages in any protest, advocacy, or dissent, or engages in any strike, lockout, or other industrial action, is not, by itself, a sufficient basis for inferring that the person—

   a. is carrying out an act for a purpose, or with an intention, specified in subsection (2); or

   b. intends to cause an outcome specified in subsection (3).[66]

From Section 22:

1. The Prime Minister may designate an entity as a terrorist entity under this section if the Prime Minister believes on reasonable grounds that the entity has knowingly carried out, or has knowingly participated in the carrying out of, 1 or more terrorist acts.
2. On or after designating an entity as a terrorist entity under this Act, the Prime Minister may designate 1 or more other entities as an associated entity under this section.
3. The Prime Minister may exercise the power given by subsection (2) only if the Prime Minister believes on reasonable grounds that the other entity—

   a. is knowingly facilitating the carrying out of 1 or more terrorist acts by, or with the participation of, the terrorist entity (for example, by financing those acts, in full or in part); or

   b. is acting on behalf of, or at the direction of,—

      i. the terrorist entity, knowing that the terrorist entity has done what is referred to in subsection (1); or

      ii. an entity designated as an associated entity under subsection (2) and paragraph (a), knowing that the associated entity is doing what is referred to in paragraph (a); or

   c. is an entity (other than an individual) that is wholly owned or effectively controlled, directly or indirectly, by the terrorist entity, or by an entity designated under subsection (2) and paragraph (a) or paragraph (b).

4. Before designating an entity as a terrorist or associated entity under this section, the Prime Minister must consult with the Attorney-General about the proposed

[66] "New Zealand Terrorism Suppression Act 2002," Parliamentary Counsel Office, 2020, link.

designation.[67]

Possible Changes to Definition of Terrorism

The prospective amendments to the Terrorism Suppression Act 2002 and the definition of a "terrorist act" are as follows:

Clause 6 amends section 5, which is the definition of terrorist act. The definition includes an act that falls within section 5(2). Some requirements, relating to purpose and intention, in section 5(2) are adjusted so that—

1. the act must be carried out for "1 or more purposes that are or include" (not for "the purpose of") advancing an ideological, political, or religious cause; and
2. the act must be carried out with the intention—
   a. to induce "fear in a population" (not "terror in a civilian population");or
   b. "to coerce" (not "to unduly compel") or to force a government or an international organization to do or abstain from doing any act.

Section 5(3), which specifies required outcomes referred to in section 5(2), is also amended so that—

1. section 5(3)(d) specifies serious interference with, or serious disruption to, "critical infrastructure" (not "an infrastructure facility"), if likely to endanger human life (and with the new term "critical infrastructure" having the meaning given to it in section 4(1) as amended by clause 5); and
2. section 5(3)(e) specifies introduction or release of a disease-bearing organism, if likely to "cause major damage to" (not "devastate") the national economy of a country.[68]

Violent Extremism

The New Zealand Security Intelligence Service (NZSIS) released new definitions of violent extremism – to include Identity-Motivated Violent Extremism or Faith-Motivated Extremism – "adapted from a helpful framework developed by our Canadian sister agency" and "makes it clear that our concern is with violent extremists and terrorists of varying ideologies."[69]

NZSIS uses the following terminology when referring to extremist ideology:

1. Faith-Motivated Violent Extremism (FMVE): promoting the use of violence to advance one's own spiritual or religious objectives.

[67] "Final Designation, New Zealand Terrorism Suppression Act 2002," Parliamentary Counsel Office, 2020, link.
[68] "Counter-Terrorism Legislation Bill, New Zealand Terrorism Suppression Act 2002," Parliamentary Counsel Office, 2020, link
[69] "Counter-Terrorism Legislation Bill," Parliamentary Counsel Office.

2. Identity-Motivated Violent Extremism (IMVE): promoting the use of violence to advance one's own perception of identity and/or denigrate others' perceived identities.
3. Politically Motivated Violent Extremism (PMVE): promoting the use of violence to achieve change to or within an existing political system.
4. Single-Issue Motivated Violent Extremism (SMVE): promoting the use of violence to achieve a desired outcome to a specific issue; and
5. White Identity Extremism (WIE): describes extremely radical ideologies and beliefs that are focused on real or perceived threats to concepts of a white or ethnic-European culture and identity.[70]

United Kingdom

According to the U.K. Crown Prosecution Service, terrorism is defined as the use or threat of action, both in and outside of the U.K., designed to influence any international government organization or to intimidate the public. It must also be for the purpose of advancing a political, religious, racial, or ideological cause.

Examples include:

1. serious violence against a person or damage to property,
2. endangering a person's life (other than that of the person committing the action),
3. creating a serious risk to the health or safety of the public or a section of the public,
4. action designed to seriously interfere with or seriously to disrupt an electronic system.

It is important to note that in order to be convicted of a terrorism offence a person doesn't actually have to commit what could be considered a terrorist attack. Planning, assisting and even collecting information on how to commit terrorist acts are all crimes under British terrorism legislation.

Terrorism is defined in section 1 of the Terrorism Act 2000 (as amended). Section 1(4) of the Terrorism Act 2000 provides that the references to action, persons, property, the public, and the government apply to the U.K. or abroad:

1. In this Act "terrorism" means the use or threat of action where—
   a. the action falls within subsection (2),
   b. the use or threat is designed to influence the government [F1 or an international

[70] 2020 Annual Report, New Zealand Security Intelligence Service (NZSIS), 2020, link.

governmental organization] or to intimidate the public or a section of the public, and

   c. the use or threat is made for the purpose of advancing a political, religious [F2, racial] or ideological cause.

2. Action falls within this subsection if it—
   a. involves serious violence against a person,
   b. involves serious damage to property,
   c. endangers a person's life, other than that of the person committing the action,
   d. creates a serious risk to the health or safety of the public or a section of the public, or
   e. is designed seriously to interfere with or seriously to disrupt an electronic system.
3. The use or threat of action falling within subsection (2) which involves the use of firearms or explosives is terrorism whether or not subsection (1)(b) is satisfied.
4. In this section—
   a. "action" includes action outside the United Kingdom,
   b. a reference to any person or to property is a reference to any person, or to property, wherever situated,
   c. a reference to the public includes a reference to the public of a country other than the United Kingdom, and
   d. "the government" means the government of the United Kingdom, of a Part of the United Kingdom or of a country other than the United Kingdom.
5. In this Act a reference to action taken for the purposes of terrorism includes a reference to action taken for the benefit of a proscribed organization.[71]

Violent Extremism

The Counter Extremism Strategy 2015 says that "Extremism is the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist."[72] This definition is also included in the U.K.'s Prevent strategy, which covers all forms of terrorism, including far-right extremism and some aspects of non-violent extremism.[73]

[71] "Terrorism: Guidance in relation to the prosecution of individuals involved in terrorism overseas," The Crown Prosecution Service, September 2019, link.
[72] "Counter-Extremism Strategy," HM Government United Kingdom, Accessed on June 1, 2021, link.
[73] "Prevent Duty Guidance for England and Wales," HM Government United Kingdom, Accessed on June 4, 2021, link.

United States of America

The United States' primary definitions for international terrorism and domestic terrorism are given in Title 18 of the United States Code.

According to 18 USC § 2331(1):

5. The term "international terrorism" means activities that—
   a. involve violent acts or acts dangerous to human life that are a violation of the criminal laws of the United States or of any State, or that would be a criminal violation if committed within the jurisdiction of the United States or of any State;
   b. appear to be intended—
      i. to intimidate or coerce a civilian population;
      ii. to influence the policy of a government by intimidation or coercion; or
      iii. to affect the conduct of a government by mass destruction, assassination, or kidnapping; and
   c. occur primarily outside the territorial jurisdiction of the United States, or transcend national boundaries in terms of the means by which they are accomplished, the persons they appear intended to intimidate or coerce, or the locale in which their perpetrators operate or seek asylum;[74]

According to 18 USC § 2331(5) and recently reiterated in the White House National Security Strategy for Countering Domestic Terrorism:

5. the term "domestic terrorism" means activities that—
   a. involve acts dangerous to human life that are a violation of the criminal laws of the United States or of any State;
   b. appear to be intended—
      i. to intimidate or coerce a civilian population;
      ii. to influence the policy of a government by intimidation or coercion; or
      iii. to affect the conduct of a government by mass destruction, assassination, or kidnapping; and
   c. occur primarily within the territorial jurisdiction of the United States;75

Title 18 of the United States Code also includes a definition for the Federal crime of terrorism in 18 USC § 2332b(g)(5), which concerns Acts of Terrorism Transcending National Boundaries. Though the United States Code does not have a statute that defines a crime of

[74] 18 U.S.C. § 2331 (2018), Retrieved July 1, 2021, link.

[75] 18 U.S.C. § 2331; "National Strategy for Countering Domestic Terrorism," White House National Security Council, June 15, 2021, link.

domestic terrorism, 51 of the 57 offenses in § 2332b(g)(5) can be used in both international and domestic terrorism cases. Additionally, cases that meet the federal statutory definition of domestic terrorism in 18 USC § 2331(5) are sometimes classified and prosecuted as federal hate crimes, or bring charges that are not referenced in § 2332b(g)(5), such as "interstate threats, firearms offenses, offenses against government employees and false statement or fraud offenses."[76]

According to 18 USC § 2332b(g)(5):

5. the term "Federal crime of terrorism" means an offense that—
    a. is calculated to influence or affect the conduct of government by intimidation or coercion, or to retaliate against government conduct; and
    b. is a violation of—
        i. section 32 (relating to destruction of aircraft or aircraft facilities), 37 (relating to violence at international airports), 81 (relating to arson within special maritime and territorial jurisdiction), 175 or 175b (relating to biological weapons), 175c (relating to variola virus), 229 (relating to chemical weapons), subsection (a), (b), (c), or (d) of section 351 (relating to congressional, cabinet, and Supreme Court assassination and kidnaping), 831 (relating to nuclear materials), 832 (relating to participation in nuclear and weapons of mass destruction threats to the United States) 842(m) or (n) (relating to plastic explosives), 844(f)(2) or (3) (relating to arson and bombing of Government property risking or causing death), 844(i) (relating to arson and bombing of property used in interstate commerce), 930(c) (relating to killing or attempted killing during an attack on a Federal facility with a dangerous weapon), 956(a)(1) (relating to conspiracy to murder, kidnap, or maim persons abroad), 1030(a)(1) (relating to protection of computers), 1030(a)(5)(A) resulting in damage as defined in 1030(c)(4)(A)(i)(II) through (VI) (relating to protection of computers), 1114 (relating to killing or attempted killing of officers and employees of the United States), 1116 (relating to murder or manslaughter of foreign officials, official guests, or internationally protected persons), 1203 (relating to hostage taking), 1361 (relating to government property or contracts), 1362 (relating to destruction of communication lines, stations, or systems), 1363 (relating to injury to buildings or property within special maritime and territorial jurisdiction of the United States), 1366(a) (relating to destruction of an energy facility), 1751(a), (b), (c), or (d) (relating to Presidential and Presidential staff assassination and kidnaping), 1992 (relating to terrorist attacks and other acts of violence against railroad carriers and against mass transportation systems on land, on water, or through the air), 2155 (relating to destruction of national defense materials, premises, or utilities), 2156 (relating to national defense material, premises, or utilities),

[76] E. Halliday and R. Hanna, "How the Federal Government Investigates and Prosecutes Domestic Terrorism," Lawfare, February 16, 2021, link.

2280 (relating to violence against maritime navigation), 2280a (relating to maritime safety), 2281 through 2281a (relating to violence against maritime fixed platforms), 2332 (relating to certain homicides and other violence against United States nationals occurring outside of the United States), 2332a (relating to use of weapons of mass destruction), 2332b (relating to acts of terrorism transcending national boundaries), 2332f (relating to bombing of public places and facilities), 2332g (relating to missile systems designed to destroy aircraft), 2332h (relating to radiological dispersal devices), 2332i (relating to acts of nuclear terrorism), 2339 (relating to harboring terrorists), 2339A (relating to providing material support to terrorists), 2339B (relating to providing material support to terrorist organizations), 2339C (relating to financing of terrorism), 2339D (relating to military-type training from a foreign terrorist organization), or 2340A (relating to torture) of this title;

ii. sections 92 (relating to prohibitions governing atomic weapons) or 236 (relating to sabotage of nuclear facilities or fuel) of the Atomic Energy Act of 1954 (42 U.S.C. 2122 or 2284);

iii. section 46502 (relating to aircraft piracy), the second sentence of section 46504 (relating to assault on a flight crew with a dangerous weapon), section 46505(b)(3) or (c) (relating to explosive or incendiary devices, or endangerment of human life by means of weapons, on aircraft), section 46506 if homicide or attempted homicide is involved (relating to application of certain criminal laws to acts on aircraft), or section 60123(b) (relating to destruction of interstate gas or hazardous liquid pipeline facility) of title 49; or

iv. section 1010A of the Controlled Substances Import and Export Act (relating to narco-terrorism).[77]

Title 22 of the United States Code provides a separate definition of terrorism, around which are built a unique definition of international terrorism and a definition of terrorist groups. These definitions primarily concern the Secretary of State's duty to send annual reports on terrorism to Congress and the Secretary of State's authority to designate an organization as a Foreign Terrorist Organization.

According to 22 U.S. Code § 2656f:

> (d)(1) the term "international terrorism" means terrorism involving citizens or the territory of more than 1 country;
>
> (d)(2)the term "terrorism" means premeditated, politically motivated violence perpetrated against noncombatant targets by subnational groups or clandestine agents;
>
> (d)(3) the term "terrorist group" means any group practicing, or which has significant

[77] 18 U.S.C. § 2332 (2018), Retrieved July 1, 2021, link.

sub-groups which practice, international terrorism;[78]

Title 31 of the Code of Federal Regulations provides an additional definition of terrorism. Section 594 of Title 31 focuses on Global Terrorism Sanctions Regulations and concerns the Office of Foreign Assets Control ("OFAC") of the U.S. Department of the Treasury.

According to 31 CFR § 594.311:

The term terrorism means an activity that:

a. Involves a violent act or an act dangerous to human life, property, or infrastructure; and
b. Appears to be intended:
    i. To intimidate or coerce a civilian population;
    ii. To influence the policy of a government by intimidation or coercion; or
    iii. To affect the conduct of a government by mass destruction, assassination, kidnapping, or hostage-taking.[79]

Violent Extremism

According to the 2017 Strategic Implementation Plan for Empowering Local Partners to Prevent Violent Extremism in the United States, violent extremists are defined as "individuals who support or commit ideologically-motivated violence to further political goals."[80]

According to the United States Government Accountability Office report, "Countering Violent Extremism: Actions Needed to Define Strategy and Assess Progress of Federal Efforts," violent extremism is "ideologically, religious, or politically motivated acts of violence... perpetrated in the United States by white supremacists, anti-government groups, and radical Islamist entities, among others."[81]

International and Regional Organizations

Given the government representatives on GIFCT's Independent Advisory Committee (IAC), the following regional organizations definitions are also included here as they are contextually and regionally relevant. International bodies and organizations tend to have definitions that are transnational and well resourced.[82]

[78] 22 U.S.C. § 2656 (2018), Retrieved July 1, 2021, link.
[79] "Global Terrorism Sanctions Regulations," 31 CFR § 594 (2020), Retrieved July 1, 2021, link.
[80] "Strategic Implementation Plan for Empowering Local Partners to Prevent Violent Extremism in the United States," Department of Homeland Security, Accessed May 9, 2021, link.
[81] "Countering Violent Extremism Actions Needed to Define Strategy and Assess Progress of Federal Efforts," United States Government Accountability Office, Report to Congressional Requesters, Accessed June 4, 2021, link.
[82] Chris Meserole and Daniel Byman, "Terrorist Definitions and Designations Lists," Global Research Network on Terrorism and Technology: Paper No. 7 (2019).

European Union

Directive (EU) 2017/541 of the European Parliament and of the Council of 15 March 2017 on combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA:

Title II -Terrorist Offences and Offences Related to a Terrorist Group

Article 3 -Terrorist offences

1. Member States shall take the necessary measures to ensure that the following intentional acts, as defined as offences under national law, which, given their nature or context, may seriously damage a country or an international organization, are defined as terrorist offences were committed with one of the aims listed in paragraph 2:

   a. attacks upon a person's life which may cause death;

   b. attacks upon the physical integrity of a person;

   c. kidnapping or hostage-taking;

   d. causing extensive destruction to a government or public facility, a transport system, an infrastructure facility, including an information system, a fixed platform located on the continental shelf, a public place or private property likely to endanger human life or result in major economic loss;

   e. seizure of aircraft, ships or other means of public or goods transport;

   f. manufacture, possession, acquisition, transport, supply or use of explosives or weapons, including chemical, biological, radiological or nuclear weapons, as well as research into, and development of, chemical, biological, radiological or nuclear weapons;

   g. release of dangerous substances, or causing fires, floods or explosions, the effect of which is to endanger human life;

   h. interfering with or disrupting the supply of water, power or any other fundamental natural resource, the effect of which is to endanger human life;

   i. illegal system interference, as referred to in Article 4 of Directive 2013/40/EU of the European Parliament and of the Council (19) in cases where Article 9(3) or point (b) or (c) of Article 9(4) of that Directive applies, and illegal data interference, as referred to in Article 5 of that Directive in cases where point (c) of Article 9(4) of that Directive applies;

   j. threatening to commit any of the acts listed in points (a) to (i).

2. The aims referred to in paragraph 1 are:

   a. seriously intimidating a population;

   b. unduly compelling a government or an international organization to perform or abstain from performing any act;

c. seriously destabilizing or destroying the fundamental political, constitutional, economic or social structures of a country or an international organization.[83]

United Nations

The United Nations (U.N.) does not have an agreed definition of terrorism but instead provides a Consolidated List which includes all individuals and entities subject to measures imposed by the Security Council.[84] The U.N. list was originally put in place as a sanctions list of entities affiliated with Al-Qaeda, the Taliban and the Islamic State, which the U.N. compiles pursuant to Resolutions 1267, 1989 and 2253.37[85]

Tech Companies

The section below focuses on GIFCT founding members and their respective definitions for terrorism and violent extremism. Just like governments, intergovernmental institutions, civil society organizations, and academics, tech companies often have slightly different definitions of terrorism, terrorist content and violent extremism.

We have included the founding members of GIFCT, who tend to have greater capacity, but for reference our Resource Guide contains the efforts of the other GIFCT member companies to counter terrorist and violent extremist activity and tools developed to combat forms of online radicalization.[86]

Facebook

According to Facebook's Community Standards on Dangerous Individuals and Organizations, terrorist organizations and terrorists are any non-state actor that:

1. Engages in, advocates, or lends substantial support to purposive and planned acts of violence,
2. Which causes or attempts to cause death, injury or serious harm to civilians, or any other person not taking direct part in the hostilities in a situation of armed conflict, and/or significant damage to property linked to death, serious injury or serious harm to civilians,
3. With the intent to coerce, intimidate and/or influence a civilian population, government, or international organization in order to achieve a political, religious,

[83] "Combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA" (Directive 2019/904), Council of the European Union, 2017, link.
[84] "United Nations Security Council Consolidated List," United Nations Security Council, Retrieved June 24, 2021, link.
[85] "United Nations Security Council Consolidated List," United Nations Security Council.
[86] "Resource Guide," GIFCT, 2021, link.

or ideological aim.[87]

Furthermore, Facebook defines hate organizations as "any association of three or more people that is organized under a name, sign, or symbol and that has an ideology, statements, or physical actions that attack individuals based on characteristics, including race, religious affiliation, nationality, ethnicity, gender, sex, sexual orientation, serious disease or disability."[88]

In terms of violating content and behaviors, Facebook does not allow:

1. Symbols that represent any of the above organizations or individuals to be shared on our platform without context that condemns or neutrally discusses the content.
2. Content that praises any of the above organizations or individuals or any acts committed by them.
3. Coordination of support for any of the above organizations or individuals or any acts committed by them.
4. Content that praises, supports, or represents events that Facebook designates as terrorist attacks, hate events, mass murders or attempted mass murders, serial murders, hate crimes and violating events.

In addition to the above definitions and clarification of violating content, Facebook also does not allow Militarized Social Movements (MSM), such as militias or groups that support and organize violent acts amid protests, or Violence-Inducing Conspiracy Networks, such as QAnon, to maintain a Page, Group, Event or Instagram profile, or to have one maintained on their behalf.[89]

Microsoft

According to the blog post titled "Microsoft's approach to terrorist content online," Microsoft recognizes that "there is no universally accepted definition of terrorist content." For its services, Microsoft considers terrorist content to be material posted by or in support of organizations included on the United Nations Security Council (UNSC) Consolidated List that "depicts graphic violence, encourages violent action, endorses a terrorist organization or its acts, or encourages people to join such groups."[90] For context, the U.N Sanctions List includes groups that the UNSC determines to be terrorist organizations.[91]

[87] "Facebook Community Standards - Dangerous Individuals and Organizations," Facebook, 2021, link.
[88] "Facebook Community Standards," Facebook.
[89] "Facebook Community Standards," Facebook.
[90] "United Nations Security Council Consolidated List," United Nations Security Council.
[91] "Microsoft's approach to terrorist content online" (blog), Microsoft, June 13, 2017, link.

Twitter

According to Twitter's Violent Organizations Policy, violent extremist groups are those that meet all of the following criteria:

1. They identify through their stated purpose, publications, or actions as an extremist group;
2. They have engaged in, or currently engage in, violence and/or the promotion of violence as a means to further their cause; and
3. They target civilians in their acts and/or promotion of violence.[92]

Other violent organizations are those that exist as "a collection of individuals with a shared purpose" and "have systematically targeted civilians with violence."[93] This policy states that users cannot "affiliate with and promote the illicit activities of a terrorist organization or violent extremist group."[94] Examples of violating content include:

1. Engaging in or promoting acts on behalf of a violent organization;
2. Recruiting for a violent organization;
3. Providing or distributing services (e.g., financial, media/propaganda) to further a violent organization's stated goals; and
4. Using the insignia or symbol of violent organizations to promote them or indicate affiliation or support.

YouTube

According to the Featured Policies: Violent Extremism in Google's Transparency Report, YouTube states that violating content under its violent extremism policies includes "material produced by government-listed foreign terrorist organizations." YouTube also does not allow terrorist organizations to use the platform for any purpose, including recruitment.

The platform also strictly prohibits content that "promotes terrorism, such as content that glorifies terrorist acts or incites violence." The platform does, however, make allowances for content shared in an "educational, documentary, scientific, or artistic context."

In addition to content produced by government-listed foreign terrorist organizations, YouTube also addresses violent extremist groups that are not government-listed foreign terrorist organizations in its "policies against posting hateful or violent or graphic content,

[92] "Twitter Help - Our policy on violent organizations," Twitter, April 2, 2021, link.
[93] "Twitter Help - Our policy on violent organizations," Twitter.
[94] "Twitter Help - Our policy on violent organizations," Twitter.

including content that's primarily intended to be shocking, sensational, or gratuitous."[95]

According to YouTube's Violent Criminal Organizations Policy:

In the context of its violent criminal organizations policy, YouTube instructs users not to post:

1. Content produced by violent criminal or terrorist organizations
2. Content praising or memorializing prominent terrorist or criminal figures in order to encourage others to carry out acts of violence
3. Content praising or justifying violent acts carried out by violent criminal or terrorist organizations
4. Content aimed at recruiting new members to violent criminal or terrorist organizations
5. Content depicting hostages or posted with the intent to solicit, threaten, or intimidate on behalf of a violent criminal or terrorist organization
6. Content that depicts the insignia, logos, or symbols of violent criminal or terrorist organizations in order to praise or promote them

YouTube clarifies that users posting "content related to terrorism or crime for an educational, documentary, scientific, or artistic purpose" should be "mindful to provide enough information in the video or audio itself so viewers understand the context."[96]

[95] "Featured Policies: Violent Extremism," Google Transparency Report, Google, Retrieved 24 June 2021, link.
[96] "Violent criminal organizations policy," YouTube Help, YouTube, May 28, 2019, link

Bibliography

18 U.S.C. § 2331 (2018). LII / Legal Information Institute. Retrieved July 1, 2021. link.

18 U.S.C. § 2332 (2018). LII / Legal Information Institute. Retrieved July 1, 2021. link.

22 U.S.C. § 2656 (2018). LII / Legal Information Institute. Retrieved July 1, 2021. link.

2020 Annual Report, New Zealand Security Intelligence Service (NZSIS). 2020. link.

"Code Pénal." France, French Government, link.

"Combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA" (Directive 2019/904). Council of the European Union. 2017. link.

"Counter-Extremism Strategy." HM Government United Kingdom. Accessed on June 1, 2021. link.

"Counter-Terrorism Legislation Bill, New Zealand Terrorism Suppression Act 2002." Parliamentary Counsel Office. 2020. link

"Countering Violent Extremism Actions Needed to Define Strategy and Assess Progress of Federal Efforts." United States Government Accountability Office, Report to Congressional Requesters, Accessed June 4, 2021. link.

"Crime of Preparation of Acts of Terrorism and Other Organized Crimes." Japan, Ministry of Justice. 2020. link.

"Facebook Community Standards - Dangerous Individuals and Organizations." Facebook. 2021. link.

"Featured Policies: Violent Extremism." Google Transparency Report, Google, Retrieved 24 June 2021. link.

"Final Designation, New Zealand Terrorism Suppression Act 2002." Parliamentary Counsel Office. 2020. link.

"Global Terrorism Sanctions Regulations." 31 CFR § 594 (2020). link.

"Governance." GIFCT. 2021. link.

Halliday, E. and R. Hanna. "How the Federal Government Investigates and Prosecutes Domestic Terrorism." Lawfare. February 16, 2021. link.

"Justice Laws Website." Legislative Services Branch. 2019. link.

Meserole, Chris and Daniel Byman. "Terrorist Definitions and Designations Lists." Global Research Network on Terrorism and Technology: Paper No. 7. 2019.

Merriam-Webster, s.v. "Terrorism." Accessed April 21. 2021. link.

Merriam-Webster, s.v. "Extremism." Accessed June 1. 2021. link.

"Microsoft's approach to terrorist content online" (blog), Microsoft, June 13. 2017, link.

"National Strategy for Countering Domestic Terrorism." White House National Security Council. June 15, 2021. link.

"New Zealand Terrorism Suppression Act 2002." Parliamentary Counsel Office. 2020. link.

Oxford University Press, s.v. "Terrorism." Accessed April 21. 2021. link.

Oxford University Press, s.v. "Extremism." Accessed April 21. 2021. link.

"Prevent Duty Guidance for England and Wales." HM Government United Kingdom. Accessed on June 4, 2021. link.

"Resource Guide." GIFCT. 2021. link.

"Strategic Implementation Plan for Empowering Local Partners to Prevent Violent Extremism in the United States." Department of Homeland Security. Accessed May 9, 2021. link.

"Terrorism: Guidance in relation to the prosecution of individuals involved in terrorism overseas." The Crown Prosecution Service. September, 2019. link.

"Threats to the security of Canada and Canadian Interests." Service, C. S. I., May 20, 2020. link.

"Twitter Help - Our policy on violent organizations." Twitter, April 2. 2021. link.

"Understanding definitions of terrorism." Briefing European Parliamentary Research Service. 2015. link

"United Nations Security Council Consolidated List." United Nations Security Council, Retrieved June 24. 2021. link.

"Violent criminal organizations policy." YouTube Help, YouTube, May 28. 2019. link.

Yamoah, N. Anti -Terrorism Act. 2008 (ACT 762). Ghana Justice. August 14, 2020. link.

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