tech against terrorism

GIFCT
Global Internet Forum to Counter Terrorism

GIFCT Technical Approaches Working Group

Gap Analysis and Recommendations for deploying technical solutions to tackle the terrorist use of the internet

July 2021

Author: **Tech Against Terrorism**

GIFCT
Global Internet Forum
to Counter Terrorism

GIFCT
Global Internet Forum to Counter Terrorism

Table of Contents

Introduction to the Gap Analysis on Technical Approaches (2021)

Objective of This Report

The objective of this report is to provide strategic guidance to tech companies, government policy makers, and solution providers in order to increase and improve investment into effective technical approaches that support platforms in tackling the terrorist use of internet services while respecting human rights. While stopping short of providing a detailed roadmap for development, this report aims to provide the overall framework for prioritization of effort given the complex nature of the terrorist threat and the challenges faced by tech companies in adapting to this threat and increased regulatory pressures from governments.

To achieve this objective this report evaluates gaps between the technical requirements of smaller tech platforms in moderating terrorist content and availability of solutions in order to inform the overall strategy of the Working Group through developing priority technical requirements, formulating a high-level roadmap, and analyzing delivery models that are appropriate for platforms of all sizes and capabilities.

Background to This Report

The findings of this assessment are drawn from meetings of the Global Internet Forum to Counterterrorism's (GIFCT) Working Group on Technical Approaches (TAWG) co-led by Facebook, the UK Home Office, and Tech Against Terrorism (TaT). The TaT team has written this report based on the suggestions of Working Group members and TaT's background understanding of terrorist and violent extremist content dissemination and removal on smaller platforms.

Scope

Platforms in scope

Platforms in scope are content-sharing platforms that are vulnerable to terrorist use of their services where there is a demonstrable need for detecting and actioning content at scale. Priority will be given to those platforms that have a high "risk over capability ratio" as identified in the Platform Prioritization Framework later in the report. This scope includes all platforms that enable the sharing of user-generated content. This includes but is not limited to file sharing, file storage, social media, archiving, link-shortening, content-pasting, email, messaging, video sharing, and blogging platforms. Due to the long tail of platforms used by terrorists, TaT estimates that at any

one point there are 250-500 platforms used by designated terrorist organizations to disseminate content.

Terrorist Content

The scope of this report is limited to content clearly associated with designated terrorist organizations and the underlying technologies designed to support platforms in managing overall removal processes for this content. TAWG stresses the need to focus not only on specific content detection technologies but also the general concept of "technical approaches" that encompasses the totality of content moderation policies, processes, and systems.

TAWG does not seek to define "terrorist or violent extremist" content (TVEC) in this report. Instead, TAWG refers to the **Group Inclusion Policy** underpinning TaT's Terrorist Content Analytics Platform (TCAP) and platforms' own guidance on terrorist organizations. TAWG assumes that TVEC refers to content that is clearly associated or affiliated with a designated terrorist organization. TAWG accepts that this framing is far from perfect given the deficiencies of government designation processes. However, TAWG believes that in the fight against terrorist use of the internet, focus is essential. Furthermore, this scope provides solid grounding in key democratic processes and the rule of law.

Technical Solutions

In this report, we assess the effectiveness and availability of technical solutions and initiatives. Technologies include software, methodologies, and models. Use-cases include content detection (hashing, metadata analysis, hash-sharing database, TCAP, and others), alternatives to content removal, crisis response and coordination, anti-recidivism (cross-platform coordination), community-based takedown requests, and content moderation workflow.

Deployment Models

Successful deployment of technical approaches in support of content moderation requires the development of policy, underlying data systems, culture, processes, and the recruitment and up-skilling of people. Each technology type will have its own dependencies. This report evaluates the most appropriate "go to market" strategies for tech type including deployment model: buy or reconfigure an existing solution (initially cheaper but high configuration costs), develop new solutions (more configurable but more expensive), or partnership (highly likelihood of consensus but risks stakeholder inertia).

Summary of recommendations

General Recommendations for Policymakers and Tech Platforms

Define success from the perspective of tech platforms, governments, civil society, and researchers.

- **Develop measurable and achievable objectives.** Without understanding and evaluating the desired end-state for stakeholders, it is difficult to prioritize efforts based on an objective assessment of feasibility. Many apparent objectives such as "eliminate terrorist content from the internet" or "remove content within 1 hour" are unrealistic given the limited support currently provided to smaller tech platforms.
- **Define a clear end-state.** Without consensus regarding the end-state, there is a risk of content moderation efforts leading to unintended consequences such as increased resilience by terrorist actors and migration to platforms where moderation and monitoring are more difficult (e.g. decentralized technologies). It is unclear the extent to which stakeholders are comfortable with pushing terrorists to ever-smaller platforms and who is then responsible for managing this risk.
- **Devise this success framework prior to investing in technical approaches.** This framework should aim to quantify the scale of the problem prior to the solution being implemented and then monitor progress against these objectives. Systems such as TCAP can be used to inform the baseline (e.g. in determining the observable quantity of terrorist content on smaller platforms and content moderation response rates from high-risk smaller platforms).

Formulate a strategy that encourages stakeholders to work together towards a common goal.

- **Devise a clear and coherent strategy** to tackle the terrorist use of the internet based on the objectives identified by "Recommendation 1 – Define Success." Without understanding the desired end-state and potential unintended consequences (e.g. platform migration, human rights violations, dilution of rule of law), online counterterrorism efforts are unlikely to be effective at scale. best practices such as analytical methods from systems thinking should be incorporated into this strategy.
- **Increase information sharing** between policymakers, the intelligence community, tech platforms, and TVEC researchers should be encouraged on a more systematic basis (legal and regulatory constraints permitting). Sensitive topics

are difficult for governments to discuss outside of the intelligence community, for instance with other governments or with tech platforms. This can either result in inertia or overlapping activities that then become difficult to de-escalate. To our knowledge, there are no systematic mechanisms to support deconfliction of outreach and engagement with tech platforms.

- **Convene a multi-sector group to share threat intelligence** and notifications of imminent takedown initiatives likely to have an impact on our collective response. At present, limited threat intelligence sharing results in a lack of coordination that terrorist actors easily exploit. The scope of a multi-sector threat intelligence group should include threat sharing, trends analysis, technical approaches, regulatory interventions, transparency, human rights impacts and legal considerations. The lack of such a mechanism has impacts on a range of efforts including attempts to tackle Terrorist Operated Websites (TOWs).
- **Identify approaches beyond content removal.** It should be stressed that while terrorist use of the internet is still prevalent on smaller platforms, in general the absolute volume of content is low. Existing initiatives such as TCAP result in more than 90% of verified terrorist content being removed within 2 weeks. While additional investment in content moderation policies, processes, and systems will certainly help accelerate the timeframe of content removal, it is unlikely to change the absolute amount of terrorist content prevalent on the internet in the medium term.
- **Focus the strategy based on need.** Analysis of TCAP URL alerts since November 2020 shows that more than 80% of all content discovered on all smaller platforms (100+ platforms) is shared on the top 20% of these platforms (22 out of 115). This suggests that a targeted approach to support platforms is likely to be most effective given the long tail of platforms being used. This also indicates that most smaller platforms are receptive to content alerts already.

Devise a prioritized roadmap based on the magnitude of the threat and likely impact of investing in technical approaches.

- **The roadmap** should consider two approaches: **1) Prioritizing platforms most in need** (see the Platform Prioritization Framework below); 2) Develop technical approaches (see below) **based on the greatest impact.** Technical approaches should be considered a broad category of enabling activities and prioritization of effort should be concentrated where the most impact can be delivered. See below for the full list of technical approaches, many of which are more focused on process and people than the development of systems.
- **Consideration should be given to the Operational Expenditure (OPEX) vs. Capital Expenditure (CAPEX)** trade-off. In many cases, the most effective technical solution is hiring an analyst or developer (i.e. OPEX) to work on a

specific task rather than trying to develop a generalized solution (i.e. CAPEX) that may still need considerable maintenance. A good example of this is hiring a team of Open Source Intelligence (OSINT) analysts (as has TaT in support of the Terrorism Content Analytics Platform) who have the expertise required to carry out ad hoc investigations on platforms.

Ensure technical solutions are considered alongside policy responses

- **Policy formulation and regulation** regarding TVEC frequently takes place without consideration for practical implementation by tech platforms. Examples of this include discussion about emergency content removal protocols without the concomitant development of tools to automate the dissemination of alerts. Even when such tools are envisaged, these tools are often late to market and focus only on a specific jurisdiction (meaning that they are of limited use to most platforms).
- **The technical approaches work should where possible be integrated with other workstreams** such as Legal and Transparency. Given ambiguous and complex legislative requirements, it is imperative that the roadmap is informed by a sound legal understanding of cross-jurisdictional legal liabilities emanating from the risk associated with data protection, data privacy, defamation, copyright, and counterterrorism legislation. This work will also be informative for platforms themselves. The technical approaches working group should commission legal analysis to inform how these risks can be mitigated rather than seeking to avoid risk altogether.

Establish a fund to finance and implement technical solutions to identify and mitigate the exploitation of smaller platforms by terrorists and violent extremists.

- **The lack of funding available to support smaller platforms represents a market failure:** the platforms most in need of support have the least economic capacity to respond effectively. To help address this we recommend establishing a fund to finance and implement technical solutions to identify and manage the removal of terrorist content on smaller platforms. Many smaller tech companies cannot afford to develop or purchase tech solutions to deploy on their services. Further, even when having access to such solutions, smaller tech companies may not have the capability or time to implement them.
- **The fund should have substantial financial resources** at its disposal sufficient to commission new tools, buy / repurpose existing technologies, and deploy a team of dedicated software engineers and analysts to support implementation. These solutions should be without cost for smaller platforms.

- **The work of the fund should be guided by the roadmap** (Recommendation 3) based on objectives laid out in Recommendation 1 and a strategy formulated as part of Recommendation 2.

Segment tech platforms by their size and capacity and ensure that this segmentation is used to prioritize effort. Often when discussing the terrorist use of the internet it is assumed that the tech sector is homogeneous in terms of its capacity and capability to tackle the terrorist use of internet services. Given this is not the case, we recommend segmenting tech platforms into the following categories: micro, small, medium, large, and very large in order to encourage differentiated strategies for each. See below for our suggested segmentation and taxonomy.

Empower GIFCT and TaT to represent the perspective of smaller platforms. Despite the strategic significance of terrorist use of smaller platforms, the perspective of smaller platforms is often overlooked by stakeholders. Often smaller platforms are inundated with requests from law enforcement for content removal or attendance at international conferences. It is unrealistic to expect micro platforms to attend the volume of meetings required of them (and in any case this would almost certainly distract them from implementing improved content moderation mechanisms).

Encourage large tech platforms to "open source" more of their content moderation technologies and share more about how they approach TVEC content moderation for the benefit of smaller platforms. This should not be limited to classifiers but instead should encompass the wide range of technical approaches recommended in this report (in particular content moderation workflows and associated tools).

Research to understand and explain the regulatory and legislative challenges faced by tech platforms in adopting improved technical approaches. One example of this is that in some jurisdictions there are apparent overlaps between requirements for data privacy and compliance for other unrelated regulatory initiatives. The result of this is an ambiguous legal framework that hinders the potential advancement of technical approaches such as hash-sharing.

- **Work with the legal approaches working group** to understand these challenges and advance the conversation.
- **Invest in mitigating, minimizing and accepting risk**. Out of an abundance of caution, larger platforms often take conservative legal positions regarding their own technical approaches and the extent to which they share data and technologies with other platforms and researchers. To be effective, the GIFCT–TaT technical approaches workstream will need to mitigate these risks both through investing significant resources to minimize risks and also through adopting a less risk-averse stance where possible.

Initial Recommendations for the Roadmap

The following section outlines our suggested initial priorities for the roadmap.

Strengthen existing initiatives that are already showing success, such as GIFCT's hash-sharing database and TaT's TCAP and its URL-sharing capability. In doing so apply the Pareto Principle to ensure that effort is concentrated where the most impact can be delivered (i.e. the top 20% of affected platforms).

Expand the GIFCT hash-sharing database. TAWG recommends that the access to the hash-sharing database be widened, and additional support provided to GIFCT members to accelerate its implementation. Additional transparency measures may encourage greater adoption and interoperability with other initiatives such as TCAP. GIFCT should consider broadening the scope of content included in the hash-sharing database (for example based on TaT's Group Inclusion Policy for TCAP).

Enhanced URL-sharing via TCAP. In the first 6 months of its operation, TCAP sent approximately 6,000 terrorist content alerts to smaller tech platforms. 90% of this content was removed within 2 weeks rising to 96% after 4 weeks. This demonstrates the effectiveness of coordinated alerting however more can be done to support emergency referral processes (e.g. GIFCT's Content Incident Protocol (CIP), the Christchurch Protocol, the EU Crisis Response Protocol, TCAP's Threat to Life Protocol).

Focus on developing content moderation workflow solutions to facilitate the enforcement of Terms of Service (ToS) by smaller platforms. At present there are few available tools to support the decision-making process by smaller platforms with the result that most rely on email or spreadsheets to track content moderation tasks. Such a system will improve the collection of data required for transparency reporting, which is currently a labor-intensive process for most smaller platforms.

Develop standalone tools that can support enhanced content moderation like Arabic script transliteration and image lookups for suspected terrorist logos (currently in the roadmap for the Knowledge Sharing Platform (KSP)).

Focus on technology supporting collaboration between platforms to ensure shared best practice for content removal both in terms of efficacy and transparency (e.g. harmonized emergency content removal processes, threat to life). See Annex 3. Summary of existing crisis protocols.

Support safety by design and regulatory risk assessments. Develop tools and approaches to support platforms in evaluating their features and tech stack to

support improved regulatory requirements for risk assessments. Small platforms are unlikely to be able to do this unaided.

Scale-up OSINT capabilities to monitor and analyze ongoing adversarial shifts by terrorist actors. As tech sector responses to terrorist content improve, so do terrorist content-sharing methodologies. It is imperative that despite very high proactive detection rates on larger platforms there is continued investment in human-led analysis of terrorist use of the internet. For instance, we are finding the increased prevalence of obscured Islamic State (IS) activity on larger platforms as a result of improved proactive detection of logos and obvious terminology. Success of the GIFCT TAWG in part should be assessed based on the extent to which terrorists and violent extremists are forced to adapt and how technical approaches are devised to anticipate this adversarial shift.

Threat Assessment - Terrorist use of the internet

Assessment and Prioritization of Current / Emerging Use of Internet Tech by Terrorists and Violent Extremists

State of Play: Terrorist Use of the Internet

Terrorists use a range of platforms spanning across technology types. In the table below we list overarching platform categories and their use-cases for terrorist content:

Platform type	Terrorist content use-case
Social media platforms	Social media platforms offer terrorists the best opportunity to reach a large external audience and to have bilateral engagements with their members, supporters, and wider populations.
Messaging apps	Messaging apps offer terrorists an easy, secure, and often free means of both internal and external communication. Most messaging apps frequently used by terrorist actors are protected by either end-to-end or client-server encryption (or give the impression of such encryption).
Alt-tech platforms	A variety of platforms have emerged in the past few years that claim to offer an alternative to larger mainstream platforms like Facebook, Twitter, and YouTube. These platforms often explicitly market themselves as "free speech" platforms, or ones that oppose the "censorship" of larger platforms. Some alt-tech platforms used blockchain-based decentralized technology. Both of these qualities are attractive to terrorists and maximizes their chances of online stability.[1] Many alt-tech platforms are Video Sharing Platforms.
Video sharing platforms (VSPs)	VSPs provide terrorists with an ideal platform through which to promote their audio-visual content. Search functions within these sites mean that content can easily be found, and file size limits are typically larger than on most other online platforms.
File hosting platforms	File hosting or pasting sites are used by terrorists to store content such as videos, images, and audio files. They are also used to aggregate information, such as lists of URLs to further content stored elsewhere.
Gaming-related platforms	Terrorists use gaming platforms to radicalize and recruit, and to propagate their ideologies through video games. They have also used chat functions within some gaming platforms to communicate, plan attacks and events, as well as stream attacks.[2]

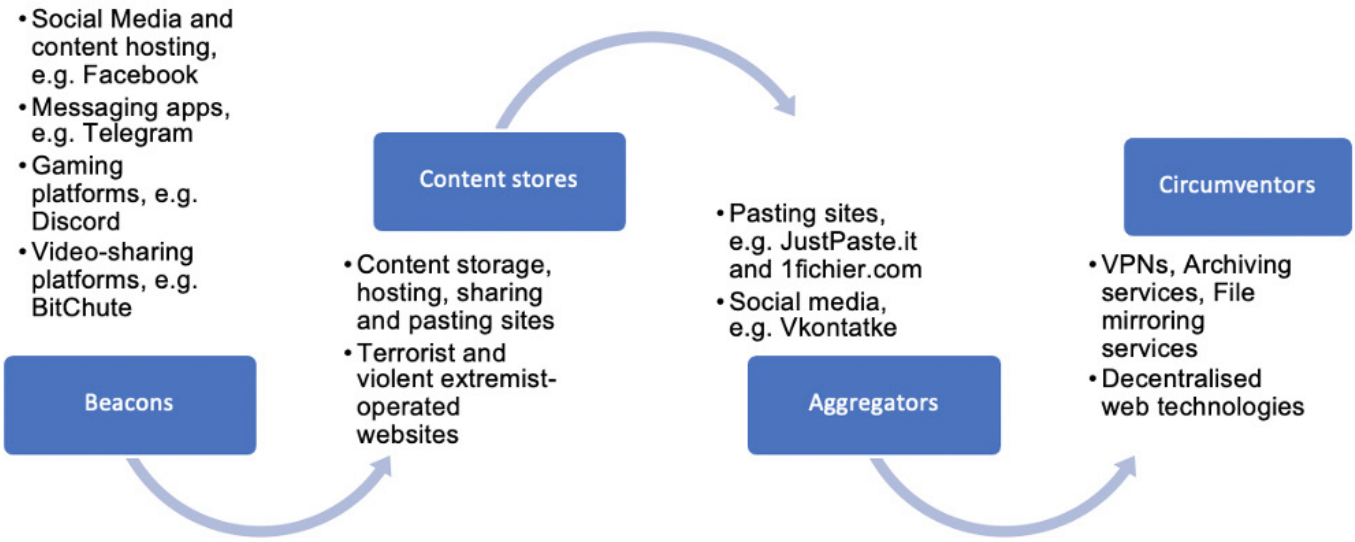
1 Alt-Tech: Far-right safe spaces online," Hope Not Hate, November 4, 2018, link
2 Linda Schlegel, "Points, Rankings and Raiding the Sorcerer's Dungeon: Top-Down and Bottom-up Gamification of Radicalisation and Extreme Violence," The Global Network on Extremism and Technology, February 17, 2020, link.

Terrorist operated websites (TOWs)	Websites that are run by terrorist groups or their supporters with the intended purpose of serving a terrorist group or network's interests. These play an important role in the online terrorist ecosystem, often acting as a centralized source of content that may have been removed from social media platforms or messaging platforms. Unlike most content on messaging apps or social media sites, content found on these sites is often indexed by search engines. TOWs play an increasingly important role in the online terrorist propaganda eco-system. At the time of writing, TaT is aware of 121 websites suspected of being operated by terrorist actors. Unlike accounts on third-party platforms like Facebook, Twitter, or Telegram, terrorists can control content on websites, as individual posts or pieces of content are not liable to content moderation. TOWs can be removed, but it involves a more drawn-out reporting process and raises more complex legal and ethical questions, based on the multi-jurisdictional nature of online domains.

In addition to the various technology types, there are generally four categories3 of online platforms used in terrorist and violent extremist content dissemination:

Category	Description
1. Beacons	Platforms used by terrorists and violent extremists to project their content to the widest audience possible. The beacon acts both as a centrally located lighthouse and a signpost to where the content can be found. Through beacons, terrorists redirect their target audience to the platforms on which content is hosted.
2. Content stores	Where terrorist content is stored, including text and audio files, as well as images and videos. These are used as online libraries of content. Terrorists and violent extremists rely on content storage platforms and pasting sites, as well as archive services.
3. Aggregators	Aggregators act as centralized databases of where content can be found online, gathering together a wide range of URLs to content hosting platforms to facilitate diffusion. If one link is taken down, terrorists can easily find an alternative to share.
4. Circumventors	Online services and platforms used to circumvent content moderation and de-platforming measures. Circumventors include VPNs, which can enable nefarious actors to access content that has been blocked in specific countries. Another example of circumventors is the use of decentralized web technologies, which avoid website takedowns.

3 The categories are based on analysis by Fisher, Prucha, and Winterbotham, to which Tech Against Terrorism adds "circumventors" as an additional category. Ali Fisher, Nico Prucha, & Emily Winterbotham, "Mapping the Jihadist Information Ecosystem: Towards the Next Generation of Disruption Capability," Global Research Network on Terrorism and Technology, Paper No. 6, 2019, link.



Networked terrorist propaganda dissemination strategies ensure that a single piece of terrorist content can be replicated a theoretically unlimited number of times. The online propaganda dissemination strategy of terrorists can be characterized as a "swarmcast"[4]; content distribution is no longer centralized but has evolved to one that involves a disparate and fluid network of hostile actors that constantly upload and redistribute content produced by designated organizations or terrorist individuals.

Online terrorist propaganda campaigns therefore affect an entire tech ecosystem of platforms. While many platforms across multiple technology types are affected, key beacon platforms and smaller file sharing and content storage platforms constitute the most important strategic threats due to their crucial role in the online information ecosystem.

Further research should be conducted to understand more about the role played by services such as mirroring sites that are essential in facilitating widespread dissemination of terrorist content. Careful focus on services with the greatest "centrality" and "connectedness" (to borrow concepts from network analysis theory) is likely to be the most efficient use of resources.

Distribution of TVEC Across Platforms

While the largest platforms tackle the greatest volume of terrorist content, the vast majority of this content is detected and actioned proactively using advanced classifiers. However, when analyzing content that is removed due to referral or alert, data from TCAP shows that the majority of content that is not addressed proactively is found on smaller platforms.

---

4 Adopted concept from Fisher, Prucha, and Winterbotham, "Mapping the Jihadist Information Ecosystem."

For smaller platforms more than 80% of TVEC is found on only 20% of smaller platforms (namely the top 22 platforms out of 115 in TCAP's dataset). This indicates that wherever possible we should prioritize those platforms that are being used most by terrorists.

Based on our research we find that terrorists look for four attributes in an online platform:

Attribute	Description
Security	Terrorists look for features offering enhanced security and privacy. End-to-end encryption is an example, alongside private or secret chats and servers.
Stability	Terrorists look for platforms where they can establish a stable presence, for example due to the platform having limited capacity, capability, or willingness to ban accounts or remove content.
Audience reach	Features that increase audience reach are attractive to terrorists because they allow for straightforward and widespread propaganda dissemination.
Usability	Terrorists favor platforms whose design and features make them user-friendly. This enables faster and more straightforward content storage and sharing. Features like search functions make groups and channels easier to find.

Terrorists aim to use platforms that have all of the above attributes, and platforms that possess most or all of these attributes are more likely to be targeted by terrorists. However, in reality few platforms meet all the criteria. Terrorists therefore will often need to assess the benefits and downsides of each platform. For example, terrorists might sacrifice security benefits if a platform can provide wide audience reach, and conversely choose a less popular platform if it allows for more stable dissemination. In summary, terrorists deliberately choose which platforms to invest time and resources into, and being aware of the characteristics they seek in platforms allows tech companies to better assess future adversarial shifts.

Adversarial Complications

Adversarial Shift Leading to Parallelization and Fragmentation of Content-sharing

Smaller platforms typically have limited capacity, resources, and capability to devise appropriate policies, processes, and systems in support of content moderation efforts.[5] This means that terrorist actors can with limited effort establish themselves on vulnerable

5 Tech Against Terrorism, "Analysis: ISIS use of smaller platforms and the DWeb to share terrorist content," April 2019, link; Ali Fisher, "Swarmcast: How Jihadist Networks Maintain a Persistent Online Presence," Perspectives on Terrorism (Vol. 9, No. 3, 2015), pp. 3-20; Fisher, Prucha and Winterbotham, "Mapping the Jihadist Information"; forthcoming Terrorist Content Analytics Platform transparency report (Tech Against Terrorism).

platforms with the result that content remains unmoderated for long enough for terrorist content and activity to become prevalent on a platform.

Tech solutions targeting terrorist content online are currently lacking in terms of anticipating and addressing the adversarial shift we see from terrorist actors online, as well as considering the role tech solutions have in influencing such a shift. There are two key components to this shift: platform migration and content moderation circumvention techniques.

Platform Migration

Terrorist organizations such as IS and al-Qaeda now depend on smaller encrypted messaging apps and file-sharing platforms due to removal activity by large platforms such as Facebook, Twitter, and Telegram. While it is unquestionably positive that terrorists struggle to keep their content on the larger platforms, it is unknown whether such campaigns have produced an improved threat situation overall. This is because the threat has dispersed across an array of predominantly smaller and micro platforms. As a result, terrorist groups have adopted content "swarming" as a key strategy to ensure content longevity online.[6]

Removal campaigns can also risk uncontrolled migration and diffusion of terrorist activity. A campaign in 2019 by a government organization to remove IS channels from Telegram, while in the short-term limited some IS use of Telegram, led to increased exploitation of smaller messaging apps that were caught entirely unprepared for sudden influxes of IS activity. The fact that much of this activity by law enforcement was conducted in secret meant that many smaller platforms were unable to mitigate migration risk.

Content Moderation Circumvention

Terrorist groups adapt their content moderation circumvention techniques specifically to avoid detection and removal from automated tech solutions. Such measures might include alternative spellings (names, accounts, hashtags), image and video manipulation, and language / imagery sanitization. Tech solutions are currently not adept at catching such techniques in an accurate manner and often require almost constant updating. At the time of writing, the TaT OSINT team is observing a significant increase in the use of large platforms by IS to share content that closely resembles conventional news stories. It is likely such activity will increase as terrorist actors are forced to adapt their content and TTPs to improved content moderation efforts.

Moderation circumvention techniques include:

---

6 Fisher, 'Swarmcast: How Jihadist Networks Maintain a Persistent Online Presence', Perspectives on Terrorism (Vol. 9, No. 3, 2015), pp. 3–20; Fisher, Prucha and Winterbotham, "Mapping the Jihadist Information Ecosystem."

Tactic	Description
Mirroring	Terrorists and violent extremists create multiple identical accounts, or simultaneously upload multiple copies of the same content via file-mirroring platforms. The aim is to overwhelm content moderation teams by creating more accounts than they are capable of moderating.
Private channels and/or servers	Terrorist organizations and groups will often respond to takedowns of public groups and channels by creating private, invite-only versions. Join links to the channel can be shared within and outside the platform.
Content editing and repurposing	Content produced by terrorist organizations is often edited and repurposed to avoid automated takedowns (for example by blocking out branding or segmenting illegal content between content that is more admissible, such as mainstream news media reporting). Pro-IS groups often blur out the logo of Amaq News, an official IS outlet, from the top right corner of video productions. On larger platforms TaT is increasingly seeing IS and AQ posting content that closely resembles legitimate news, occasionally interspersing news content with content designed to promote radicalization.
Language amendments	Terrorists avoid keyword detection by tech platforms by amending terms and phrases that may already be on the radar of content moderation teams. They may insert spaces and underscores in the middle of key phrases or change their language entirely. TAWG has seen Telegram channels containing Arabic IS content, for example, change their titles to Mandarin.
Rhetoric dilution	Terrorists intentionally dilute their rhetoric to avoid violating content standards and de-platforming.
Misrepresentation	Terrorists share propaganda content, framing it as "journalistic" content, and that they "do not endorse" the material being shared.
Outlinking	Terrorists post content via third-party platform outlinks to avoid detection from content moderation teams (particularly when the linked content would be picked up by automated detection systems if it were posted in-app).
Archiving	Terrorists use web archiving services to create backed-up copies of content that has been uploaded to file-sharing platforms.

Terrorists invest in their online operational security (OPSEC). Dedicated OPSEC groups provide terrorist groups and violent extremist networks with advice and guidance on remaining anonymous and operating securely online.[7] For example, IS-affiliated Electronic Horizons Foundation serves as the tech knowledge sharing arm of IS since 2016, having started as an "IP support desk" for IS supporters. The site focuses exclusively on sharing online security tips rather than terrorist content. It has become a well-known OPSEC portal in the terrorist online space in the past few years via its dedicated website.

7 Michael Loedenthal, "Digital Resiliency and OPSEC strategies amongst clandestine networks," Global Network on

Prioritization of Technical Need for Each Tech Type (E.g. Content Detection, Crisis Response, Anti-recidivism, User Referrals, Moderation Workflow Based on a Prioritization Framework)

Platform Prioritization Framework

TAWG proposes adopting a prioritization framework that considers the quantity of terrorist content on a platform and the platform's capacity to detect and action this content. The intention of this framework is to guide the selection of technical approaches and focus on the needs of platforms facing the greatest risk of terrorist use. In most cases, this framework implies smaller platforms have a significantly greater need for support. However, large platforms still present a risk given their scale and increasingly sophisticated adversarial responses by terrorist actors. In summary, the suggested framework includes:

1. Absolute quantity of terrorist content shared by users;
2. Prevalence of terrorist content;
3. Proactivity of content identification and action (including redress); and
4. Timeliness of response for referred content that has not been proactively detected.

According to public transparency reports, in 2020 Facebook removed 33.3m pieces of content for violation of its Dangerous Organizations: Terrorism and Organised Hate policies. Defining "prevalence" as the proportion of content views for removed content versus total content viewed over the same time, this implies a prevalence of around 0.06%.[8] This means that out of every 10,000 views of content on Facebook, no more than 6 of those views contained content that violated their policy. An alternate definition of prevalence calculates the proportion of TVEC removed from a platform versus content uploaded over the same period. This metric is usually more appropriate for smaller platforms given limited content consumption data typically available.

By contrast, smaller platforms analyzed by TaT suggest that some have a terrorist content prevalence of between 10%-50% (whether defined on the basis of views or uploads).

Effective prioritization of platform support requires an accurate understanding of these key metrics. With TCAP now fully operational, TaT can provide its own estimates of observable TVEC quantity, prevalence, and timeliness of actioning (however this should be supplemented by the platform's own transparency reports where available). More

Extremism and Technology Insight, September 10, 2020, https://gnet-research.org/2020/09/10/digital-resiliency-and-opsec-strategies-amongst-clandestine-networks/.

8 See https://transparency.fb.com/data/community-standards-enforcement/dangerous-organizations/facebook

research is required to develop a robust approach to estimating actual TVEC on platforms given the limitation of only being able to analyze content that is directly observed.

Proactive Content Removal

By absolute number, most terrorist content posted online is removed proactively using content detection algorithms hosted on the larger platforms. For Facebook, of the total amount of TVEC removed, 99.6% was removed proactively (33.1m in 2020). Other large platforms such as YouTube, TikTok, and Twitter have similarly high levels (90% and above) of proactive terrorist and violent extremist content detection, corresponding to extremely low levels of prevalence.

Nevertheless, despite the successes of algorithmic content removals on the large platforms, their sheer scale means for platforms such as Facebook approximately 140,000 pieces of terrorist content were still removed due to referral in 2020. By contrast, most smaller platforms typically do not have the capability to proactively remove content.

Technical Gap Analysis

Drivers of Need for Support by Tech Platforms

Myth of small platform non-compliance

Sensationalist media coverage of the terrorist use of the internet gives the false impression that terrorists are running amok online. By contrast, most large platforms automatically remove 95%+ of terrorist content and most smaller platforms respond to takedown requests within hours of being alerted. Data from TaT's TCAP shows that since November 2020 96% of URLs pointing to verified terrorist content was removed by smaller platforms.

Nevertheless, there is a small minority of platforms (around 5% of all smaller platforms by our estimates) who are reluctant to engage with TVEC content moderation requests. Some of these are "alt-tech" platforms specifically designed to cater to content that has been pushed off more mainstream platforms. Others are likely directly associated with designated terrorist organizations (e.g. TOWs or niche Video Sharing Platforms designed with extreme far-right users in mind).

Pressures Faced by Tech Platforms

The vast majority of smaller platforms are keen to improve their content moderation efforts. Other than civic duty, the main driver of change for smaller platforms is increased regulatory and legislative pressure.

Regulatory and legislative measures
Tech platforms are facing increasing legal requirements from government regulation that directly and indirectly concerns technical approaches. TaT's analysis of emerging global regulation[9] shows that countries like India and Pakistan are directly encouraging tech platforms to increase their reliance on automated content removal. Further, jurisdictions like Australia, the EU, Germany, India, Pakistan, Turkey, and Indonesia currently have legislation in place that compels tech companies to remove content within a specified timeframe,[10] which will inevitably encourage tech platforms to introduce automated removal solutions. The newly passed EU terrorist content online regulation can also compel tech platforms to introduce "specific measures" which could include the introduction of automated tech approaches. The EU's draft Digital Services Act would give authorities the power to audit tech platforms' use of automated tooling. It is unclear at this stage whether this would include platforms' use of externally developed solutions.

9 See the Online Regulation Series: link.
10 Ranging from 1 to 48 hours.

Assessment of the Challenges Faced by Platforms in Scaling Content Moderation

In this section, we consider the major technical gaps that constrain effective moderation of terrorist content by tech platforms. TAWG categorizes these gaps based on constraints (Capability, Capacity, Coordination) and the functional areas within a tech platform (Policies, Processes, People, Systems).

Platform Segmentation by Size

In our day-to-day work with content sharing platforms, TAWG uses the following segmentation to guide our mentorship work based on proxies of revenue and scale such as total number of employees. TAWG suggests that where possible technical approaches are developed with this segmentation in mind given the significant disparity in capability and capacity between micro, small, medium, and large platforms.

Estimate of platform capacity and activity according to size				
	Micro	**Small**	**Medium**	**Large**
Total number of employees	0-2	1 – 15	15 -75	75+
Total number of moderators	0	1	2-4	5+
In-house legal advisors	0	0	1-2	3+
Community guidelines		Yes	Yes	Yes
Internal recording of takedown requests	Yes	Yes	Yes	Yes
Transparency reporting according to size of platform	Yes	Yes	Yes	Yes

Constraints to Adoption of Improved Technical Approaches

Limited capacity. While larger platforms have a pool of resources to draw from, some tech platforms (especially smaller platforms) have severely limited capacity to develop improved content moderation tools and engage with stakeholders. Most of the "micro platforms" exploited by terrorists are so small that they have no outside investment and limited monetization. As a result, many smaller platforms cannot afford to hire specialists to devise content moderation policies, map business processes underlying content moderation, run hash-matching algorithms, develop automated solutions, or

buy externally developed technical solutions.

Limited bandwidth to engage with external stakeholders (for instance at conferences) means that smaller platforms are underrepresented in broader discussions, and so their perspective is often overlooked by policymakers. See Recommendation 7.

Limited capability. It is often assumed by policymakers that all technology is alike and that smaller platforms have the capability to develop content moderation tools and classifiers. Many of the high-risk platforms being used by terrorists are straightforward products (blogs, pasting websites, link sharing systems, mirroring sites), have relatively simple tech stacks, and are built using existing frameworks or development environments. This means that there is a wide variety of technical capabilities among platform founders. By contrast, many of the technical solutions designed for content moderation require advanced knowledge of cloud engineering, Python libraries, database management, and in some cases advanced data science.

Where possible TAWG recommends the prioritization of services that alert smaller platforms to the existence of likely terrorist content and help them manage content moderation processes rather than trying to re-engineer platforms. There are currently several automated or semi-automated alert, notification, and/or referral mechanisms that flag terrorist content to smaller tech platforms.[11] There are also known cases in which third parties have developed classifiers and automated detection tools to support platforms in discovering terrorist content online.[12] However, no centralized resource to facilitate the creation and implementation of automated solutions to support smaller tech companies exists.

Coordination challenges. The limited capacity of smaller platforms not only impacts their ability to implement technical solutions, but it also constrains their ability to engage with external stakeholders. As a result of poorly coordinated activity from stakeholders, many smaller platforms become easily overwhelmed by duplicated content removal requests from a wide range of law enforcement authorities. The result of this is that some platforms remove content without due process or they stop engaging with content moderation requests entirely.

Gaps With Platform Policies, Processes, People, Systems

Smaller tech companies sometimes struggle to introduce the policies, processes, and systems required to facilitate the effective implementation of technical solutions. Many of

11 This includes so-called Internet Referral Units (IRUs), which have been set up by the EU, the United Kingdom, France, and the Netherlands.
12 One example includes the video classifier developed by Faculty AI on behalf of the UK Home Office, which aimed to identify IS videos. See more: link

the smallest platforms exploited by terrorist groups might only have rudimentary policies prohibiting terrorist use of their services (or none at all). Smaller platforms also struggle to put in place effective processes to manage moderation of such activity, including the systems they use to facilitate that process.[13] Such policies, processes, and systems will need to be in place before platforms implement automated technical solutions to ensure that they are useful. Much of the joint effort of GIFCT and TaT is focused on mentorship which is a free service designed to support platforms in introducing a robust baseline for policy and processes.

Overview of Technical Approaches

The following provides a high-level overview of the main technical approaches considered in this report. We recommend focusing on all stages of product development:

Technical Approach		Description
1. Safety by design		Anticipating adversarial use by terrorists, red-teaming functionality within product management process; devising effective risk assessments to inform development in advance of product / feature launch and to support regulatory requirements
2. Technical removal capability		Ensuring platform has the technical ability to remove content and record when doing so (contributes to Transparency Reporting below)
3. Content alerting / referral / reporting		Reporting of URLs
	3a) Internal	Users reporting a URL for ToS violation from within the app
	3b) External URL-sharing	Trusted 3rd party sharing URLs of concern
	3c) Emergency referral processes	Law enforcement sharing URLs related to emergency referral process
4. Proactive content detection		
	4a) Ad hoc investigations	Analyst-led investigation: searching for users and content based on using analytical techniques such as network analysis and understanding terrorist tactics, techniques, procedures (TTPs), terrorist content adversarial shift, terrorist content libraries
	4b) Hash-matching	Use of cryptographic hashes to match samples of content

13 Confidential interviews with TaT member and mentee companies.

	4c) Content classifiers	Automated detection of likely terrorist content based on prior similar content or inclusion of high-risk attributes such as terrorist logos, terminology, and imagery
	4d) Metadata analysis	Analysis of metadata associated with groups / channels (e.g. group size, velocity of growth, international dispersal of users) to flag unusual activity on the platform for further investigation
5. Transparency reporting		Development of transparency reporting tools to reduce time required to generate these and improve detail
6. Content moderation		Supporting the content moderation workflow
	6a) Providing context for content moderators	Providing resources to content moderators (see the KSP) to improve content moderation decisions by providing additional context (library of logos, terrorist group TTPs, terminology, or similar prior content) based on content detected on other platforms
	6b) Workflow management	Supporting content moderators in managing moderation decisions and facilitating more efficient removal of content (i.e. not using email but bespoke tools to support the decision workflow required)
	6c) Redress and appeal	Ensuring that the content moderation workflow supports the ability for users to request redress and appeal for content moderation decisions they may disagree with
	6d) Anti-recidivism	Prevent violating users or content returning to the platform or easily migrating elsewhere
	6e) Moderation tools	Specific tools to help moderators with common tasks such as translating text in images (Arabic transliteration), matching logos against known lists, etc. (some already supported by the KSP)
7. Content intervention mechanisms		Mechanisms that platforms use to limit the spread of terrorist content (i.e. not just removing the content or users associated with it)
	7a) Alternatives to content removal	In addition to content removal, experiment and test the success of alternatives such as placing warning notices and password-protecting content in order to ensure that future adversarial shifts do not render content removal approaches redundant
	7b) Positive interventions	Build features into the product to minimize the risk of radicalization (e.g. high-risk content recommendation algorithms) and implement positive interventions such as counter-narrative content / off-ramping links

Details of each technical approach as well as the associated requirement, availability, complexity, cost, and capacity, priority, Rationale and implementation recommendations are listed in Annex 6.

Evaluation of Gap Between Platform Needs and Available Solutions

Barriers to Adoption

In the following section, TAWG compares the need of each technical approach with the feasibility of its implementation. In follow-on work from this report, TAWG recommends more comprehensive research is undertaken to support the product management process and analyzes the feasibility of each solution according to:

1. **Availability.** Does the solution already exist? Can this be repurposed from elsewhere or does this need to be built?
2. **Complexity.** How complex is the solution to implement? How does the impact the delivery model? (See Deployment Considerations below)
3. **Cost.** Given availability and complexity, how much will this solution cost? Is this a matter of one-off investment or ongoing support?
4. **Capacity.** How much capacity is required from the small platform to engage with a given technical approach?

As a result, we recommend first focussing on technical solutions that address the basics (such as whether a platform can technically remove content). While it may seem absurd to suggest that platforms cannot remove content, some decentralized platforms had not thought this through prior to launch and were forced to develop hasty mechanisms to remove content (often borrowing code used to remove content based on copyright violations).

Addressing the expectation gap

Tech solutions, and especially automated solutions, are effective in scaling up otherwise time-consuming manual processes. However, to do this effectively and accurately, they need to be supported by meaningful policies and processes. Tech solutions alone cannot therefore address many underlying challenges that many smaller platforms face around building out effective and human rights compliant moderation enforcement practices.[14]

Addressing "AI hype"

There is a gap with regards to what many stakeholders expect tech solutions to be able to do and what they are effective at. What is needed is to clarify exactly what role

---

14 See for example Alexander Stamos, "Prepared written testimony and statement for the record before the US House of Representatives Committee on Homeland Security, Subcommittee on Intelligence and Counterterrorism on 'Artificial Intelligence and Counterterrorism: Possibilities and Limitations'," June 25, 2019, link.

automated solutions can and should play based on current capability. In doing so we should move away from flawed notions, including that "artificial intelligence" will be able to effectively tackle online terrorist content without sufficient human guidance. Instead, we should build consensus around what exactly tasks and workstreams technology should support, including content identification and removal but also moderation processes and workflows.

Deployment Recommendations

Priority Technical Requirements and Prerequisites (People, Policy, Processes, Underlying Data Systems, Technology Integration Requirements)

In the following section, we evaluate the most appropriate "go to market" strategies for tech types (including deployment model): buy or reconfigure an existing solution (initially cheaper but high configuration costs), develop new solutions (more configurable but more expensive), partnership (highly likelihood of consensus but risks stakeholder inertia).

At present, the most significant gap for smaller platforms is in supporting internal content moderation management and workflow systems. It is vital that smaller platforms are supported in implementing more efficient moderation mechanisms, processes, and systems as a primary step, instead of attempting to implement sophisticated classifiers and automated solutions that platforms do not have the policies, processes, and systems yet in place. This is particularly relevant for the production of transparency reports which depends on the upstream moderation processes and systems.[15]

The following should be considered to support tech platforms in implementing technical approaches:

1. **Knowledge sharing.** Developing resources (for example on TaT's KSP) to facilitate improved content moderation efforts by tech platforms. Resources to include videos, manuals, and interactive tools.
2. **Documentation.** The development of comprehensive documentation (e.g. in Confluence) to support more effective implementation of code.
3. **Training.** Delivery of in-person and virtual training (to include the major technical disciplines required for the implementation of technical approaches such as cloud engineering (AWS or similar), machine learning, Python development, etc.). A recommendation is that GIFCT / TaT seek discounted training courses from major providers such as AWS, Google Cloud Platform, Microsoft Azure, etc.
4. **Product management.** The management of developers and product development according to agile principles and based on user (platform) needs.
5. **Development.** The integration or development of existing or bespoke tools using

15 For instance, a number of tech platforms investigated by TaT on the question of transparency reports explained that they do not or did not used to report on terrorist content specifically due to the lack of a dedicated reporting category in the reporting processes and ensuing content moderation workflow.

web-based tech stacks.

6. **Policy and legal analysis.** Analysis of emerging government regulations and the related requirements (or enablers) for tech platforms.
7. **Data science.** Deployment of advanced statistics and machine learning capabilities to develop predictive models operating on structured and unstructured datasets (e.g. image recognition algorithms and classifiers).
8. **OSINT.** Open source intelligence: the capability to analyze and anticipate terrorist use of platforms and understand relevant TTPs.

Balancing Immediate vs. Medium-Term Impact in Supporting Platforms

While it is critical to invest in technical approaches that will scale up and prove resilient in the medium term, it is also important that platforms that are currently high-risk are supported expeditiously. With this in mind, we recommend applying the Platform Prioritization Framework and approaching the most vulnerable platforms as soon as possible.

With this in mind and based on feedback from platforms we have drawn up a list of suggested urgent use-cases for technical approaches.

Requirement	Recommendation
Audio classifier for IS and AQ content	Commission development of audio classifier and open API for access by smaller platforms
Hash-matching against known terrorist audio	Expand GIFCT hash-sharing to include audio
Removal of content based on keyword searches of obvious TVEC	Conduct 3-4 weeks of intensive OSINT on Platform C to identify content and share findings to support improved future automated moderation
Alert platform to known TVEC channels	Onboard platform to TCAP and pass content to GIFCT for hash-sharing

Implementing and Integrating Systems in a Cloud Environment

The complexity and cost of deploying solutions largely depends on the extent to which the tool / technology is designed to be directly integrated within the platform's tech stack, linked to it via an API, or used in parallel as a standalone tool. Wherever possible we recommend avoiding direct integration with smaller platforms given the myriad of ethical, legal, and technical complications of doing so. In particular, TAWG would caution against third parties having access to smaller platforms' tech stacks out of concern for user privacy and the risk of indirectly encouraging extra-judicious government access to

content on smaller platforms.

Repurposing Existing Solutions for Content Moderation

Most of the major cloud providers such as AWS, Microsoft Azure, and Google Cloud Platform have already developed suites of content moderation services; however, most of these are extremely complex to implement. It is likely that most commercial content moderation services have been developed based on some of these underlying systems. "Building from scratch" is therefore not necessary; however, integration / cloud engineering requirements are likely to be substantial and expensive. More research should be conducted to understand the feasibility of using existing components already available on commercial services such as AWS. It would also be worthwhile understanding how big platforms such as Facebook, YouTube, and Twitter already architect their solutions in case there is an opportunity for "open sourcing" what they have already developed.

Legal Considerations for Implementing Technical Approaches

Developing technical solutions requires in-depth consideration of legal risks and compliance strategies. In developing TCAP, TaT underwent extensive legal review[1] to identify risks and mitigation strategies. In addition to counterterrorism legislation and the legal risks of collecting large amounts of terrorist content, general risks associated with developing technical solutions are (among others): data protection, misuse of private information, defamation, malicious falsehoods, copyright infringement, and breach of confidence.

Rather than adopting a "zero risk" attitude to technical approaches, we recommend investing in appropriate cross-jurisdictional legal analysis to inform how best to minimize and mitigate risks associated with improved technical approaches. Without this approach TaT would not have been able to launch TCAP.

Identification of Opportunities to Increase Technical Collaboration With GIFCT

GIFCT's hash-sharing database and its efforts in partnership with TaT to mentor smaller tech platforms have already proven successful in building the foundations for a more comprehensive cross-industry response to terrorist use of the internet. The following are suggested opportunities for enhanced technical cooperation between GIFCT companies:

1. Encourage more transparency from the larger GIFCT companies regarding their own content moderation technologies and encourage code to be "open sourced" for the benefit of GIFCT and TaT members.

2. Expand the GIFCT hash-sharing database to include URLs (connected with TCAP) and a broader set of terrorist content (e.g. related to extreme far-right groups).
3. Improve interoperability of the GIFCT hash-sharing database and TCAP to improve archiving capability and introduce the ability to cross-reference hashes using TCAP.

Recognizing the Importance of Ensuring That Technical Approaches Support and Do Not Undermine Human Rights

There are several well-reported risks associated with using automated data-driven solutions to counter terrorist use of the internet.[16] Most of this concerns the error rates and false positives that such systems might flag, largely as a result of automated solutions not being able to account for context or nuance. Such risks include:

1. Negative freedom of speech impact by accidentally removing legitimate speech content under counterterrorism policies. Some studies suggest that such error rates predominantly affect minority groups.
2. Unwarranted surveillance.
3. Lack of transparency and accountability in the development process, which hinders external reviewers to interrogate the solution.
4. Accidental deletion of digital evidence content under counterterrorism policies, much of which is crucial in terrorism and war crime trials.

In our assessment, challenges associated with human rights and content moderation automation are predominantly the result of developers paying insufficient regard to these risks, or not having access to accurate training data or guidance from subject matter and human rights experts. Furthermore, in some cases there seems to be little coordination with external stakeholders, including subject matter experts and civil society, before such solutions are brought to market.[17] More resources on the ethical risks associated with automated tools are shared in Annex 5.

16 Dia Kayyali, "Vital Human Rights Evidence in Syria is Disappearing From YouTube," WITNESS, August 2017, link; Joint Report of Electronic Frontier Foundation, Witness, and Syrian Archive, "Caught in the Net: The Impact of 'Extremist' Speech Regulations on Human Rights Content," May 30, 2019, link.

17 For more information about how Tech Against Terrorism addresses such concerns in developing the TCAP, see link.

Annex

Annex 1: Features attractive to terrorist groups for internal and external communications

Characteristic	Features attractive for internal communications	Features attractive for external communications
Security	• Private chats • Closed servers and forums (access granted depending on contact with or approval from administrators) • End-to-end encryption • Self-destruct messages • Password-protection • Minimal details required on registration, such as telephone number • Invite-only access • Screenshot alerts • Easy account deletion/data erasure • Assurance by tech platform that user details will not be passed onto authorities	• Minimal details required on registration • Assurance by tech platform that user details will not be passed onto authorities • Ability to hide sign-up details on user profiles, such as email address or telephone number
Stability	• Little content moderation, due to either limited capability or willingness by platform to remove terrorist content • No content moderation possible (for example because of E2EE) • Decentralized content distribution, making content removal difficult or impossible	• Ability to easily create multiple mirror accounts or groups/ channels
Audience reach	• Voice memos • Voice and video calls • Little or no forward limits for messages	• Widely available and used by a significant proportion of the global population • Searchable public groups and profiles • Ability for content to be shared or forwarded widely and easily and/ or go “viral” • Large group or channel size limits • Easily shareable join links
Usability	• Secure and expansive file storage capability • Easy account set-up • Low bandwidth required to function • App works on range of device types	• Free • User-friendly interface, requires little to no technical ability to use • Low bandwidth required to function • Supports range of multimedia types • Large file size limit

Annex 2: Proactive content removal statistics for major platforms

Platform	Reporting timeframe	Total amount of terrorist content removed[18]	Removed proactively
Facebook			
	Q2 2018	9,371,800	99.7%
	Q3 2018	3,078,300	99.3%
	Q4 2018	4,880,400	99.6%
	Q1 2019	8,109,800	98.9%
	Q2 2019	5,829,200	98.8%
	Q3 2019	5,122,000	98.5%
	Q4 2019	7,524,000	99.0%
	Q1 2020	6,243,300	99.1%
	Q2 2020	8,665,200	99.6%
	Q3 2020	9,670,900	99.7%
	Q4 2020	8,582,800	99.8%
	Q1 2021	8,964,000	99.6%
YouTube			
	Q4 2018	49,618	71%
	Q1 2019	89,968	77%
	Q2 2019	74,655	87%
	Q3 2019	90,035	93%
	Q4 2019	80,687	91%
	Q1 2020	258,908	93%
	Q2 2020	921,783	95%
	Q3 2020	200,642	94%
	Q4 2020	200,642	94%
	Q1 2021	82,553	95%
Twitter			

18 Based on available figures specifying content removed under counterterrorism policies.

	Q2 2018	85,243	94%
	Q1 2019	64,231	74%
	Q2 2019	83,413	87%
	Q1 2020	184,123	91%
Microsoft			
	Q1-Q2 2020	2,642	99.7%
	Q3-Q4 2020	2,436	99.1%
TikTok			
	Q1 2020[19]	806,241	86.9%
	Q2 2020	232,370	96.4%

19 Terrorist and violent extremist content is reported on as part of anti-hate speech enforcement.

Annex 3. Summary of Existing Crisis Protocols

EU Crisis Protocol

In 2019, the EU Internet Forum committed to creating a crisis protocol to prevent viral spread of terrorist material in the immediate aftermath of a terrorist attack. The protocol was created in response to the Christchurch Call to Action. The protocol is a voluntary mechanism by which governments and tech companies commit to identify, notify, and share information about terrorist content that risks becoming viral. All contributing parties have an assigned point of contact. In the event of a potential "crisis" (defined as "where terrorist and violent extremist content spreads online rapidly"), the protocol asks contributing partners to take target (attack location, number of platforms associated content is found on, attack type, and victims affected) and impact (content virality, reproducibility, and resilience) into account to assess whether an event meets the crisis threshold. Based on that assessment, parties notify and share information to prevent content virality. Post-crisis reports are also produced and shared between contributing partners.

GIFCT Incident Response Framework

In response to the 2019 Christchurch terrorist attack, GIFCT developed and announced their own Content Incident Protocol. This was updated in 2021 to become their Incident Response Framework. This framework comprises three response levels: Content Incident Protocol (CIP), Content Incident (CI), Incident (I). These levels are designed to provide streamlined communication, information-sharing and situational awareness between GIFCT members. When a CIP or CI is activated GIFCT member companies become aware of, quickly assess, and act on potential content circulating online resulting from a real-world terrorism or violent extremist event. All hashes of perpetrator produced content is shared through the GIFCT hash database with other GIFCT member platforms. Member companies also remain in continuous communication throughout the incident.

Christchurch Call Shared Crisis Response Protocol

One of the priorities of the Christchurch Call to Action was to develop a shared crisis protocol to allow for improved information sharing between government and tech companies in the event of a crisis. While there is not much publicly available information on how the protocol works, in December 2019 the protocol was reviewed by representatives of government, the tech industry, and civil society and the recently published Christchurch Call to Action Crisis Response Workstream lays out the prioritise for developing this over the next year.

Terrorist Content Analytics Platform (TCAP): Threat to Life Protocol

TaT has developed a protocol for its TCAP to ensure that relevant law enforcement

authorities are alerted in a genuine threat to life situation. The Threat to Life Protocol (TTLP) is based on the UK Home Office and National Police Chiefs' Council definition of a threat to life and is based on UK Terrorism legislation, including the Terrorism Act 2000. Risks are assessed as either low, medium, or high. In the event of high risk, TaT will in the first instance alert UK authorities.

Annex 4. Sample of Technical Approaches and Developer for Each

The below table is based on publicly available information about the technology or specific products listed. This does not cover products offered by companies specialized in developing tools for content moderation (e.g. ActiveFence), for which there is limited public information. In future work in support of the GIFCT TAWG, we recommend the collation of a more comprehensive overview of available solutions, as this was not in scope for this initial report.

Application area	Technology / Specific product	Developer	Description	Availability
COMO workflow / decision-making	Microsoft Azure – Review Tool	Microsoft	A front-end dashboard for the Microsoft Azur Content Moderator Tool, supporting decision-making by combining machine learning and human moderators review to facilitate content moderation workflow. Available for different types of content, the review tool is to be used in conjunction with other automated solutions to facilitate the management of the moderation process. The tools allow users to use default or custom workflows to sort and track content and assign content to review teams.[20] This tool can also be used to automate the creation of human reviews when moderation API results come in.	Medium – costed
COMO workflow / decision-making	Terrorist Content Analytics Platform	TaT	Tech companies with access to the platform have an overview of the URLs alerted to them and which ones are still online. They also have the possibility to sort and filter the URLs to facilitate their moderation workflow.	High – free

20 Depending on the content and moderators' experience levels.

COMO decision-making	Scene-understanding[21]	No specific developer of automated tools using this technology has been identified	Scene-understanding is a technology that aims to build a human-like vision for machines, using AI to understand the content of an image or video content to support automated and accurate decision-making. This technology goes beyond the detection of visual features to "extract information related to the physical world which is meaningful for human operators."[22]	Low – technology is still in development phase and costly to develop or implement
COMO decision-making	Sentiment analysis[23]	No specific developer of automated tools using this technology has been identified	Similar to scene-understanding, AI is used to understand the context of text content, to identifying tones, (e.g. sarcasm or anger), opinions, and emotions, and ultimately to understand the actual mood and feeling of the writer. This goes beyond key words detection and text analysis by allowing the machine to understand the tone and context of a text similarly to how a human would do it. This technology can support moderators with the review process and keeping track of content under review, and ultimately with decision-making as the machine learns.	Low – technology is still in development phase and costly to develop or implement
Safeguarding content moderators' mental health	CleanView	ActiveFence	A browser add-on for Chrome designed for first responders for the internet" regularly exposed to "horrific content online. CleanView automatically blurs and grey images detected video content. It also allows for users to schedule regular "mindfulness breaks" (including breathing and mindfulness exercises).	High – free

21 See link.
22 See link.
23 See link.

Annex 5. Resources on Ethical Considerations and Risks Associated With Using Automated Data-driven Solutions

Ethical and Human Rights Risks in the use of Automated Tools in Content Moderation

No amount of "AI" in content moderation will solve filtering's prior-restraint problem: Emma Llansó, 23.04.2020.

This piece discusses how the technical realities of content filtering stack up against the protections for freedom of expression in international human rights law.

Unboxing Artificial Intelligence: 10 steps to protect Human Rights: Council of Europe Commissioner for Human Rights, May 2019.

This recommendation on AI and human rights provides guidance to Member States on the ways in which the negative impact of AI systems on human rights can be prevented or mitigated, focusing on 10 key areas of action.

Artificial Intelligence & Human Rights: Opportunities & Risks: Filippo Raso, Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz, and Levin Kim. Berkman Klein Center for Internet & Society at Harvard University. 25.09.2018.

This report explores the human rights impacts of AI technologies. It highlights the risks that AI, algorithms, machine learning, and related technologies may pose to human rights, while also recognizing the opportunities these technologies present to enhance the enjoyment of the rights enshrined in the Universal Declaration of Human Rights. The report draws heavily on the United Nations Guiding Principles on Business and Human Rights ("Guiding Principles") to propose a framework for identifying, mitigating, and remedying the human rights risks posed by AI.

Human Rights in the Age of Artificial Intelligence: AccessNow, November 2018. AccessNow conducts this preliminary study to scope the potential range of AI and human rights issues that may be raised today or in the near future.

Exploring the Human Rights Dimensions of Artificial Intelligence and Online Content Moderation at the IGF: Miru Lee, Association for Progressive Communications, 10.01.2020.

Discusses one of the agendas of the 14th annual meeting of the Internet Governance Forum (IGF): AI and human rights. According to the article, the threat to human rights and privacy because of AI was one of the main themes at the IGF. In particular, many

panels discussed AI ethics and principles to protect human rights.

Use of AI in Online Content Moderation: Cambridge Consultants on behalf of Ofcom, 2019.

This report examines the capabilities of AI technologies in meeting the challenges of moderating online content and how improvements are likely to enhance those capabilities over the next five years.

The impact of algorithms for online content filtering or moderation: European Parliament Policy Department for Citizens' Rights and Constitutional Affairs Directorate-General for Internal Policies, September 2020.

This study, commissioned at the request of the JURI Committee, addresses automated filtering of online content. The report introduces automated filtering as an aspect of moderation of user-generated materials. It presents the filtering technologies that are currently deployed to address different kinds of media, such as text, images, or videos. It discusses the main critical issues under the present legal framework and makes proposals for regulation in the context of a future EU Digital Services Act.

Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence: Niva Elkin-Koren, 29.07.2020.

This paper discusses content moderation by AI while mentioning the hashing techniques used by GIFCT and TaT. It then analyzes how using AI systems to govern speech raises serious concerns from a social welfare perspective.

Artificial Intelligence, Content Moderation, and Freedom of Expression: Emma Llansó, Joris van Hoboken, Paddy Leerssen, Jaron Harambam. Transatlantic Working Group. 26.02.2020.

This report focuses on content moderation and the use of automated systems for detecting and evaluating content at scale. It discusses content curation and questions about the role of recommendation algorithms in amplifying hate speech, violent extremism, and disinformation. For both content moderation and content curation, the paper explores the use of AI and other forms of automation. In particular, it focuses on their use in the fight against hate speech, violent extremism, and disinformation. Within this report, the authors highlight issues of AI tools and risks to freedom of expression.

Facebook's Most Recent Transparency Report Demonstrates the Pitfalls of Automated Content Moderation: Svea Windwehr, Jillian C. York. EFF. 08.10.2020.

This piece discusses automated content moderation's risk to freedom of expression

online, particularly looking at Facebook and Instagram.

The Rise of Content Cartels: Evelyn Douek, 07.05.2020.

This paper traces the origin and spread of content cartels. It examines the impulses behind demands for greater cooperation and the ways in which such cooperation can be beneficial. It further explores the failures of the current arrangements and the threats they pose to free speech. GIFCT's hash-sharing database is additionally mentioned.

Algorithmic content moderation: Technical and political challenges in the automation of platform governance: Robert Gorwa, Reuben Binns, Christian Katzenbach, 28.02.2020.

This article provides a technical primer on how algorithmic moderation works, examines some of the existing automated tools used by major platforms to handle copyright infringement, terrorism, and toxic speech, and identifies key political and ethical issues for these systems as the reliance on them grows.

Automated Moderation Must be Temporary, Transparent and Easily Appealable: Jillian C. York, Corynne McSherry. EFF. 02.04.2020.

The article recognizes that automated technology does not work at scale as it struggles to read nuance in speech the way humans can (and for some languages it barely works at all). It further notes that the use of automation results in numerous wrongful takedowns. The article stresses how automated moderation must therefore be temporary, transparent, and easily appealable.

The Limitations of Automated Tools in Content Moderation: New America.

This section of New America's "Everything in Moderation" series provides a more detailed discussion of the limitations of automated tools used for content moderation.

Promoting Fairness, Accountability, and Transparency Around Automated Content Moderation Practices: New America

In this section of "Everything in Moderation," New America provides a set of recommendations for developers, policymakers, and researchers to consider in order to promote greater fairness, accountability, and transparency around algorithmic decision-making in this space.

Ethical and Human Rights Risks in the use of Automated Tools in Content Moderation related to T/VE and Counterterrorism

Caught in the Net: The Impact of "Extremist" Speech Regulations on Human Rights Content: Abdul Rahman Al Jaloud, Hadi Al Khatib, Jeff Deutch, Dia Kayyali, and Jillian C. York, EFF (A joint publication from the Electronic Frontier Foundation, Syrian Archive, and Witness), 30.05.2019.

The report discusses how the reality of faulty content moderation must be addressed in ongoing efforts to address extremist content. It provides examples of blunt measures affecting marginalized users.

One Database to Rule them All: Svea Windwehr, Jillian C York, VOX-POL, 04.11.2020.

This article outlines concerns about GIFCT's harsh-sharing database. The concerns include reliance on automated solutions to moderate content leading to incorrectly removing legal speech.

Erasing History: YouTube's Deletion of Syria War Videos Concerns Human Rights Groups

This piece discusses how thousands of videos, some of which offer crucial evidence of war crimes, have been deleted via YouTube's algorithms. In particular, it sheds light on the hundreds of thousand videos of Syrian war atrocities that were removed by YouTube.

Civil Society Letter to European Parliament on Terrorism Database: AccessNow, 07.02.2019.

This open letter, from civil society organizations to the European Parliament, criticizes (regarding the Terrorist Content Regulation debate) the blind faith in a database to flag "terrorist content." Among the concerns are how filters are unable to understand the context and therefore are error-prone and notes the pervasive online monitoring on disadvantaged and marginalized individuals.

Joint Letter to EU Parliament: Vote Against Proposed Terrorist Content Online Regulation: Human Rights Watch, 25.03.2021.

In this letter to the EU Parliament, the limitations of automated content moderation tools in regards to terrorist content online are discussed.

Global Internet Forum to Counter Terrorism Transparency Report Raises More Questions Than Answers: Angel Diaz, Brennan Center, 25.09.2019.

This piece assesses GIFCT's first transparency report and discusses the concerns about the negative impacts its hash-sharing database poses on freedom of expression.

The flaws in the content moderation system: The Middle East case study: Eliza Campbell, Spandana Singh, Middle East Institute, 17.11.2020.

This piece discusses the limitations of content moderation automated tools. It stresses how, when it comes to moderation categories of content with more fluid delineations (such as extremist propaganda and hate speech), developing tools that can detect or remove this content with accuracy is extremely challenging. It sheds light on how automated tools for content moderation impact the Middle East and social media users there in particular.

YouTube AI deletes war crime videos as 'extremist material': Alex MacDonald, Middle East Eye, 13.08.2017.

This article discusses YouTube facing criticism after a new AI program monitoring "extremist" content began flagging and removing masses of videos and blocking channels that document war crimes in the Middle East.

Artificial Intelligence and Countering Violent Extremism: A Primer: Marie Schroeter, GNET, October 2020.

This report analyzes the ability of AI applications to contribute to countering radicalization. Mapping the possibilities and limitations of this technology in its various forms, the report aims to support decision makers and experts navigate the noise, leading to informed decisions unswayed by the current hype.

Annex 6: Details of technical approaches, prioritization and implementation recommendations

Technical approach	Small platform requirement	Availability	Complexity	Cost	Capacity	Priority (1 = High)	Rationale for prioritization	Type	Implementation recommendation
1. Safety by design									
	High	N/A	Low	Low	High	7	Anticipation of terrorist use of a platform is important (and will soon be required in some jurisdictions such as the UK) for new platforms and new features. However, priority should be on supporting existing platforms for the time being. When new regulations come into force this workstream will become more critical for platforms of all sizes.	Knowledge Sharing, Training, Development	Work with platforms to devise content risk evaluation tools to support risk management and imminent regulatory requirements (e.g. from the UK's Online Safety Bill and Ofcom's likely risk assessment stipulations).
2. Technical removal capability									
	Very high	N/A	Moderate	Low	Moderate	1	Discussion of content moderation is irrelevant if the platform cannot technically remove content.	Knowledge Sharing, Training	Share best practice with platforms to encourage common baseline for removal capability (not a difficult request, but will engage sensitive engagement especially with alt-tech platforms).
3. Content alerting / referral / reporting									

	3a) Internal	Low	N/A				9	Most smaller platforms are unlikely to have high adoption of internal referral systems given low userbase and high prevalence of use by terrorists. Internal referral systems are better suited to larger platforms with lower prevalence of TVEC.	Documentation, Development, OSINT	Scale up efforts by GIFCT and TaT / TCAP to share alerts with platforms. Encourage other stakeholders to contribute and share URLs including referral rationale and taxonomy. Improve interoperability of existing systems and increase transparency to minimize risk of content cartels emerging. Archive alerted content to improve future content moderation decisions and ensure there is an audit trial for future evaluation.

	3b) External URL-sharing	High	High	Low	Low	Low	2	External URL alerts / referrals from trusted sources require limited resources from the smaller platform and are timely. The downside is that they require investment in a tool like TCAP and rely on discovering content on smaller platforms which may become more difficult as adversarial shift continues.	Policy analysis, Documentation, Development	Ensure that emergency referral processes can link up to existing URL-sharing mechanisms such as TCAP and GIFCT's CIP. Also consider the need to alert competent authorities to content that indicates an imminent threat to life or otherwise could provide critical evidentiary value to investigations.
	3c) Emergency referral processes	Moderate	Low	Low	Low	Low	6	Similar to external URL-sharing from a technical perspective; however various protocols are not yet sufficiently established to easily map to supporting systems.		
4. Proactive content detection										
	4a) Ad hoc investigations		N/A	High	High	High	10	Searching for content based on keywords / known phrases can be effective; however, platforms typically do not have the capacity. Instead, more likely to be effective is empowering third-party researchers to help locate content on the platform if there is an external search capability.		

	4b) Hash-matching	Moderate	Moderate	Moderate	Low	High	4	Like URL-sharing, hash-matching is a fast and accurate way to tackle the majority of content found on smaller platforms. The downside is that it does require some technical knowledge and legal agreements to be in place given the sensitivity of data.	Documentation, Development, Data science	Invest in efforts to build new hash-matching approaches for a wider variety of content types (e.g. audio, video, PDFs) as well as widening the scope of terrorist organizations included by existing approaches.
	4c) Content classifiers	Low	Very high	Very high	High	High	10	Extremely complex and vulnerable to adversarial shift and therefore mostly irrelevant for deployment to smaller platforms.		
	4d) Metadata analysis	Low	High	High	Moderate	Moderate	12	Not suitable for all platforms and requires sophisticated understanding of terrorist TTPs on a specific platform (potentially combined with 4a) Ad hoc investigations).		
5. Transparency reporting										
		High	Low	Moderate	Moderate	High	11	Implementing effective content moderation policies, processes, and systems are required before transparency reports; however, these should be devised with transparency in mind.		
6. Content moderation										

	6a) Providing context for content moderators	Moderate	High	Low	Low	Low	6	Small platforms are unlikely to be well-versed in terrorist logos and terminology. Tools to help provide context are likely to improve the quality of content moderation efforts and encourage small platforms to implement robust standards.	Product management, development	Investigate buying or building content moderation workflow solutions for tech platforms.
	6b) Workflow management	High	Low	Very high	High	Low	3	Most smaller platforms do not have a workflow tool in place to support content moderation efforts. This should be a high priority given the importance of having a system to support the underlying content moderation policies and processes.	Product management, development, data science	Develop or commission specific tools to augment (but not replace) content moderation decisions (e.g. Arabic script, logo matching against databases, keyword searches against known terrorist terminology).
	6c) Redress and appeal	Low	Low	Low	Moderate	Low	8	This first requires fundamental content moderation systems to be in place.		
	6d) Anti-recidivism	High	Low	High	High	High	13	Potentially effective for smaller platforms; however, first requires underlying systems to be in place.		
	6e) Moderation tools	High	Moderate	High	High	Low	5	Moderation tools such as Arabic script transliteration and translation are vital to prevent over-removal of content based purely on language rather than whether content is assessed to be TVEC.		
7. Content intervention mechanisms										

	7a) Alternatives to content removal	Moderate	Low	Low	Low	High	14	Given the high prevalence of TVEC on some smaller platforms, content removal is likely the most appropriate solution.		
	7b) Positive interventions	Low	High	Low	Low	Low	15	Most appropriate for platforms with very large audience reach; however, more experimentation with positive interventions on smaller platforms should be encouraged once the fundamental technical approaches are put in place first.		

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