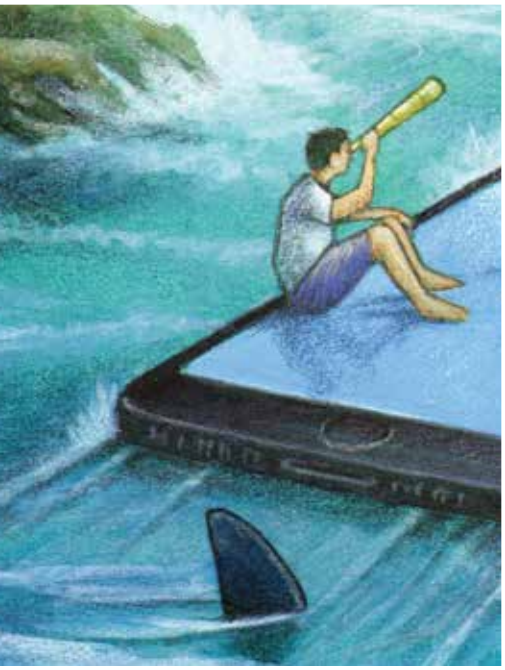
One in a Series of Working Papers from the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression



Combating Terrorist-Related Content Through AI and Information Sharing

Brittan Heller
The Carr Center for Human Rights Policy,
Harvard University

April 26, 2019

TRANSATLANTIC WORKING GROUP

The Transatlantic Working Group Papers Series

Co-Chairs Reports

Co-Chairs Reports from TWG's Three Sessions: Ditchley Park, Santa Monica, and Bellagio.

Freedom of Expression and Intermediary Liability

Freedom of Expression: A Comparative Summary of United States and European Law
B. Heller & J. van Hoboken, May 3, 2019.

Design Principles for Intermediary Liability Laws
J. van Hoboken & D. Keller, October 8, 2019.

Existing Legislative Initiatives

An Analysis of Germany's NetzDG Law
H. Tworek & P. Leerssen, April 15, 2019.

The Proposed EU Terrorism Content Regulation: Analysis and Recommendations with Respect to Freedom of Expression Implications
J. van Hoboken, May 3, 2019.

Combating Terrorist-Related Content Through AI and Information Sharing
B. Heller, April 26, 2019.

The European Commission's Code of Conduct for Countering Illegal Hate Speech Online: An Analysis of Freedom of Expression Implications
B. Bukovská, May 7, 2019.

The EU Code of Practice on Disinformation: The Difficulty of Regulating a Nebulous Problem
P.H. Chase, August 29, 2019.

A Cycle of Censorship: The UK White Paper on Online Harms and the Dangers of Regulating Disinformation
P. Pomerantsev, October 1, 2019.

U.S. Initiatives to Counter Harmful Speech and Disinformation on Social Media
A. Shahbaz, June 11, 2019.

ABC Framework to Address Disinformation

Actors, Behaviors, Content: A Disinformation ABC: Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses
C. François, September 20, 2019.

Transparency and Accountability Solutions

Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry
M. MacCarthy, February 12, 2020.

Dispute Resolution and Content Moderation: Fair, Accountable, Independent, Transparent, and Effective
H. Tworek, R. Ó Fathaigh, L. Bruggeman & C. Tenove, January 14, 2020.

Algorithms and Artificial Intelligence

An Examination of the Algorithmic Accountability Act of 2019
M. MacCarthy, October 24, 2019.

Artificial Intelligence, Content Moderation, and Freedom of Expression
E. Llansó, J. van Hoboken, P. Leerssen & J. Harambam, February 26, 2020.

www.annenbergpublicpolicycenter.org/twg

A project of the Annenberg Public Policy Center of the University of Pennsylvania in partnership with
The Annenberg Foundation Trust at Sunnylands and the Institute for Information Law of the University of Amsterdam,
with support from the Embassy of the Kingdom of the Netherlands in Washington, D.C.

TRANSATLANTIC WORKING GROUP

Combating Terrorist-Related Content Through AI and Information Sharing†

Brittan Heller, The Carr Center for Human Rights Policy, Harvard University[1]

April 26, 2019

Contents

Introduction 1
What is the GIFCT Database? 2
Is the GIFCT Database Effective? 3
Technical limitations and threats to free speech 4
Circumvention and subversion by extremist groups 5
Case study: Christchurch shooting incident 5
Conclusion 6
Notes 7

Introduction

In its first meeting at Ditchley Park, UK, in March, the Transatlantic Working Group (TWG) focused on the tech industry's efforts to address hate speech and extremist content while still respecting freedom of expression. One effort we examined was a mechanism to combat online extremism through private-to-private information sharing efforts, the Global Internet Forum to Counter Terrorism (GIFCT) and its industry-only hash-sharing database. The following analysis served as a basis for our discussions, and now has been revised to incorporate crucial insights from those discussions. It explains the background behind the GIFCT database, which often operates with secrecy, and analyses its implications for freedom of expression. This document also highlights important considerations for industry and policy makers about applying private industry-based information sharing as a method to address controversial, dangerous, or illegal online content.

---

† One in a series: A working paper of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression. Read about the TWG: https://www.ivir.nl/twg/.

What is the GIFCT Database?

In December 2016, Google, Facebook, Twitter and Microsoft announced an industry-led initiative to disrupt the terrorist exploitation of their services. This led to the June 2017 formation of the GIFCT. Its objective is to "substantially disrupt terrorists' ability to promote terrorism, disseminate violent extremist propaganda, and exploit or glorify real-world acts of violence."[2] Functionally, the GIFCT is a private enterprise to address a public harm; it is run by tech companies for the mutual benefit of tech companies.

Not much is publicly known about how the GIFCT operates. This paper inquires how private information sharing is structured and how it is technically implemented. The GIFCT's work is organized into three pillars: joint tech innovation, research, and knowledge sharing. Information-sharing efforts are housed under the GIFCT's joint tech innovation pillar, to focus on building shared technology for use within the tech industry to prevent and disrupt the spread of terrorist content online. These efforts have resulted in a common industry database of "hashes" — unique digital fingerprints — for violent terrorist imagery or terrorist recruitment videos or images that the member companies have removed from their services. When pro-terrorist content is identified and removed by one GIFCT member, the content's hash is shared with the other participating companies to enable them to identify and block the content on their own platforms.

The GIFCT's definition of terrorism is not fixed and is drawn from guidance given by the United Nations. Each member company defines and captures what qualifies as "terrorism related content" under its own terms of service. According to GIFCT's nonprofit partner, Tech Against Terrorism, the approach of defining terrorism is challenging:

> [Tech Against Terrorism] acknowledges that there is no universal definition of terrorism. In fact, one of our observations when engaging with tech companies is that they struggle with moderating content on their sites due to this uncertainty. Moreover, it is sometimes difficult to define whether a video is part of terrorist propaganda, or whether it is an important piece of news that sheds light on human right abuses. When tech companies fail to make this distinction they are often criticized, but the fact is that there is no regulating body providing clear guidelines to companies whose platforms and audiences span the entire world. Tech Against Terrorism advocates for more coherence on this matter, and therefore suggests a global normative approach rather than an ad hoc approach from single governments. We recommend companies to consult the Consolidated United Nations Security Council Sanctions list, as it provides the best framework to the international consensus on individuals and groups defined as a terrorist. Having said that, we note the absence of certain groups in that list and particularly far-right terror groups. Therefore, companies should also consult the proscribed groups and individuals' list in the specific region and/or country where the content is flagged.[3]

According to representatives affiliated with GIFCT who were interviewed for this research, public authorities do not directly interface with the shared industry hash database by design. Administratively,

the database is maintained and run by one of the four main GIFCT member companies. Governments or intergovernmental organizations reportedly do not have access to the roster of what content is indexed in the database. Efforts to enforce laws related to terrorism-related content would go through individual member companies' legal teams, but this would involve directly requesting information related to individual pieces of content through preexisting legal processes — without regard to whether or not the content is listed in the shared database.

While law enforcement or governments can theoretically come to companies with content they claim is terrorism-related, this would not necessarily mean it would be indexed by the GIFCT. The hash-sharing consortium member companies individually designate the particular content that should be flagged, tagged, or removed in accordance with their terms of service, and not against legal constraints.

With this background in mind, policy makers concerned with transparency may have concerns about the structure of the GIFCT and its relationship to public authorities. Industry claims that private information sharing has increased their capacity to respond to terrorism-related content quickly because they do not need to duplicate efforts to identify the "worst of the worst" type of information, like terrorist recruitment videos or images of graphic violence like beheadings. However, there is no external auditing of the database. Hash sharing is a closed effort, occurring outside public oversight. This raises concerns for the freedom-of-expression rights of individuals whose content mistakenly may have been flagged or accounts erroneously removed. Further, in the GIFCT database context, there is no right to any appeal for content removals. Much of the flagged content disappears from the platforms before it is even posted, making it challenging to even know if content has been removed in error. Without access to this information, there are concerns about accountability for tech companies who may be overzealous in enforcement.

Privatized efforts to deal with content that is likely to be illegal also implicate related concerns about interfacing with public authorities. Policy makers can look to a similar industry-based hash-sharing effort for child sexual exploitation issues, run by the National Center for Missing and Exploited Children (NCMEC), which was touted as having similar types of benefits for combating illegal child pornography. However, courts in the United States are still debating whether or not this structure – where private actors shared content marked as child pornography with public authorities – created a special relationship with law enforcement. If so, should that collaboration result in NCMEC qualifying as a "government entity or agent" and thus warrant Constitutional protections for the accused?[4] While courts have not yet decided, the issue of mixing private and public online enforcement mechanisms still raises concerns for digital rights. In the current regulatory landscape, where policy makers seek to designate types of controversial or harmful content as illegal, the GIFCT should provide a model to examine the limitations and challenges of looking to private companies to perform some functions that are traditionally the purview of law enforcement.

Is the GIFCT Database Effective?

The tech industry has treated this collaboration as a success that results in greater efficiency in online policy enforcement and decreases in online terrorist content. As of 2018, there were 13 hash-sharing

consortium members and the database contained more than 100,000 hashes.[5] Another 70 companies reportedly discussed joining in 2018.[6] Consortium member companies used the GIFCT hashes to identify and remove matching content – videos and images – that violated their respective policies or, in some cases, blocked terrorist content from being posted.[7]

Early data indicate that GIFCT's hash-sharing efforts are working, if content removal is the metric of success. Statistics presented by the GIFCT seem to reinforce its claim that the database is focused on a small but significant slice of content, accounting for the worst type of terrorist-related content. Between 2015 and 2017, Twitter reports having suspended over 1.2 million terrorist accounts.[8] In the second half of 2017, YouTube removed 150,000 videos for violent extremism, and over 10,000 in Q3 2018.[9] Nearly half of these were removed within two hours of upload.

Technical limitations and threats to free speech

But policy makers should recognize that challenges still remain as to the other implications of these efforts. First, there is a staggering amount of online content. Over five billion people are online. As of 2018, every single minute at least 510,000 comments and 136,000 photos were shared on Facebook, 350,000 tweets posted on Twitter, and 500 hours of video uploaded to YouTube.[10] This has increased exponentially and will only continue to grow. Given this immense number of postings, policy makers should understand that the hash-sharing database only affects a sliver of information available on the internet, so statements about impact should be contextualized.

Additionally, the GIFCT's standards do not address freedom of expression-related concerns, especially considering the problem of potential overreach when combined with technical limitations. Given the quantity of information, companies extensively rely on AI to manage identification and removal efforts. Facebook uses image matching to prevent users from uploading a photo or video that matches another photo or video that has previously been identified as terrorist-related content. YouTube has reported that 98% of the videos that it removes for violent extremism are flagged by machine-learning algorithms.[11]

However, it is important that policy makers understand that machine-learning algorithms cannot be expected to identify terrorist content with 100% accuracy. It is difficult to know how accurate these methods are in practice, since so little information is published about them. But even a 99.5% accuracy rate would create false positives affecting millions of people. Some content will be wrongly identified as "terrorist" and blocked or removed. One apparent victim of overreach is the Syrian Archive, a nonprofit aimed at documenting evidence of war crimes in the Syrian conflict. Reports from June 2018 show that its content was repeatedly removed from YouTube, leading to widespread and sometimes permanent losses of what might be crucial evidence of war crimes. According to Wired magazine, Google has taken down 123,229 of the 1,177,394 videos that Syrian Archive backed up in 2012-2018.[12] "I think we have already lost a lot of content since YouTube started using machine learning in 2017," said Hadi Al-Khatib, founder of the Syrian Archive. "There will be a big impact on the Syrian accountability process if we aren't able to retrieve it."[13]

Circumvention and subversion by extremist groups

Research about the scope and locations of online extremists' content should also factor into policy makers' efforts to evaluate the GIFCT. Consider the "whack-a-mole" dilemma, wherein companies participating in these efforts may not be the places where terrorists convene online. As a result of scrutiny from tech giants, many online extremists have migrated to the so-called dark web, to alternative gaming-based platforms like Discord, or to fringe platforms like 4chan, 8chan, and Gab for their use as communications channels.[14] These forums have been reported to have few if any restrictions on hate speech, disinformation, and other types of conduct that have led to offline violence. Online extremists have also found enforcement on major platforms to be irregular, with some platforms being more permissive than others. Also, as mentioned in the GIFCT's statement on the definition of terrorism, the database does not always capture more contextual, country-specific threat patterns and risks. In other words, there may be a distortion in the categorization of identified content – mostly ISIS-related, and less domestic extremist or white supremacist-related content – which does not match the risk profile when policy makers consider online radicalization in their countries.

In addition, policy makers need to acknowledge the technological sophistication of some extremists. Groups leveraging online content to commit offline harms are frequently early adopters of tactics to circumvent technologically oriented limitations. In response to disruption by Twitter, supporters of ISIS have tried to circumvent content blocking technology by "outlinking," spreading content through using links to other platforms. Sites often outlinked include justpaste.it (a new member of the GIFCT), sendvid.com, and archive.org. This appears to be a deliberate strategy to exploit the limited resources

Case study: Christchurch shooting incident

The technical limitations of hash-sharing technology were clearly demonstrated during recent extremist violence, accompanied by a media proliferation strategy. In the wake of the March 2019 televised attack on a mosque in Christchurch, New Zealand, the gunman's live video of the shooting circulated around the world. According to YouTube, "The volume of related videos uploaded to YouTube in the first 24 hours was unprecedented both in scale and speed, at times as fast as a new upload every second."[15] Facebook summarized the scope and scale of the attack and its online coverage:

> The video was viewed fewer than 200 times during the live broadcast. No users reported the video during the live broadcast. Including the views during the live broadcast, the video was viewed about 4,000 times in total before being removed from Facebook. Before we were alerted to the video, a user on 8chan posted a link to a copy of the video on a file-sharing site. The first user report on the original video came in 29 minutes after the video started, and 12 minutes after the live broadcast ended. In the first 24 hours, we removed more than 1.2 million videos of the attack at upload, which were therefore prevented from being seen

> on our services. Approximately 300,000 additional copies were removed after they were posted.[16]

Hash-sharing efforts failed in this instance for several reasons. Initial images did not match closely enough to any images already in the database. The shooter's first-person perspective captured clean shots to his victims. These images would not be bloody, and gore is what AI filters are often trained to identify in looking for the worst type of content. There was not enough similar preexisting content in the database to allow the machine learning to match mass shooting-related content.

Additionally, the platforms presume cooperative users who will proactively flag the worst extremist content, which is then queued to be hashed and thus prevent subsequent downloads. With a sympathetic audience waiting to spread the content, the Christchurch video was not reported until almost a half-hour after the live video began.

The Christchurch video showed the limitations of hash-sharing efforts, given the problem of virality. Mass coordination by a group of bad actors aimed to distribute copies of the video to as many people as possible through social networks, video-sharing sites, and file-sharing sites. These individuals collaborated to continually edit, upload, and create new versions of the video. The multiple versions were designed to thwart hash-sharing efforts and stymie filters looking for original versions of the content. A day after the shooting, Facebook had over 800 slightly modified duplicates of the video in its hash-sharing database.[17]

Adding to the problem was a wider population who distributed the video and unwittingly made it harder to match copies. Facebook described how "[s]ome people may have seen the video on a computer or TV, filmed that with a phone and sent it to a friend. Still others may have watched the video on their computer, recorded their screen and passed that on. Websites and pages, eager to get attention from people seeking out the video, re-cut and re-recorded the video into various formats."[18] Legitimate news outlets shared the video as well, both online and in broadcasts.[19] The variety of formats undermined existing hashing technology, and did so in a way that may have exemplified the tension inherent in examining freedom of expression and hash-sharing technology.

Conclusion

From our review, it is clear that private information sharing plays a constructive role as one tool in the toolbox for combating violent terrorist content online. It is not a panacea. But in its present form, private information sharing could be improved in order to provide better protections for freedom of expression.

For private information-sharing efforts like the GIFCT to be grounded in freedom of expression, the tech industry should adopt the following safeguards:

- Prioritize transparency;
- Develop mechanisms for increased accountability for its work, including civil-society oversight for any information-sharing models that implicate digital rights;

- Implement a right to appeal for errant content removal.

Policy makers, for their part, should take the following into consideration:

- Consider the scope and scale of internet content and extremist online activity to evaluate the full impact of the GIFCT database;
- Be aware that interactions of public authorities with private information-sharing efforts may implicate not only freedom of expression, but also other fundamental rights;
- Be aware of the technical limitations of this technology – as exemplified by deletions of online evidence by YouTube and deliberate exploitation by extremist groups of social media's hash-sharing systems;
- In particular, understand vulnerabilities emerging from hash sharing's reliance on artificial intelligence and assumptions embedded in its user design-based interfaces.

It remains to be seen what part the GIFCT collaboration will play in viable solutions to online extremism, and if it adequately protects users' ability to express themselves freely and safely on online platforms. At best, policy makers should consider it to be only a part of a multifaceted solution, given the concerns related to freedom of expression, the technical limitations of hash sharing, and evolving techniques by extremist groups to subvert indexing efforts like the GIFCT. There may be promising alternative uses of the database, like sharing images identified with viral deception. Of course, these will have the same limitations as terrorism-related content. Critically evaluating the GIFCT through the lens of freedom of expression will help policy makers and governments to protect the fundamental rights of their citizens.

Notes

[1] Brittan Heller (https://carrcenter.hks.harvard.edu/people/brittan-heller) is a technology and human rights fellow at the Carr Center for Human Rights Policy at the Harvard Kennedy School. She works at the intersection of technology, human

[2] https://www.gifct.org/about/

[3] https://www.techagainstterrorism.org/about/faq/

[4] United States v. Ackerman, No. 14-3265 (10th Cir. 2016).

[5] Ask.fm, Cloudinary, Facebook, Google, Instagram, Justpaste.it, LinkedIn, Microsoft, Oath, Reddit, Snap, Twitter and Yellow. Source: https://www.gifct.org/partners/

[6] https://www.bbc.com/news/technology-44408463

[7] The GIFCT provides more than just hash sharing to combat terrorism-related content. Under its "knowledge sharing" initiatives, small- and medium-sized tech companies are trained on strategies to address online extremism via partnerships with an affiliated non-profit, Tech Against Terrorism. It has worked with more than 100 tech companies on four continents. The GIFCT provides toolkits for startups to prevent terrorist exploitation of their platforms and services. It has also convened forums in Europe, the Asia Pacific region, and Silicon Valley for companies, civil society groups, and governments to share experiences and get suggestions for further efforts. These efforts are seen as complementary, but separate from the hash-sharing database.

[8] https://techcrunch.com/2018/04/05/twitter-transparency-report-12/
[9] https://transparencyreport.google.com/youtube-policy/featured-policies/violent-extremism?hl=en_GB
[10] https://www.dsayce.com/social-media/tweets-day/
https://www.omnicoreagency.com/youtube-statistics/
https://zephoria.com/top-15-valuable-facebook-statistics/
[11] https://www.gifct.org/about/
[12] https://www.wired.co.uk/article/chemical-weapons-in-syria-youtube-algorithm-delete-video
[13] *Id.*
[14]https://slate.com/technology/2018/10/discord-safe-space-white-supremacists.html
https://mashable.com/2017/08/17/alt-right-free-speech-online-network-gab/
[15] https://twitter.com/YouTubeInsider/status/1107645354361741312
[16] https://newsroom.fb.com/news/2019/03/update-on-new-zealand/
[17] https://newsroom.fb.com/news/2019/03/technical-update-on-new-zealand/
[18] *Id.*
[19] https://www.theatlantic.com/technology/archive/2019/03/facebook-youtube-new-zealand-tragedy-video/585418/