Recommendations on Tackling Extremist Content Online

Background

The Counter Extremism Project (CEP) is a not-for-profit, non-partisan, international advocacy organisation formed to combat the growing threat from extremist ideologies. CEP along with government agencies around the world, has been calling on technology companies, and social media platforms in particular, to rein in the ability of extremists to recruit, radicalize, plan, and execute attacks. As such, the organisation has worked to identify and report extremist content on the Internet and social media platforms like Twitter, Facebook, and YouTube since its launch in 2014.

Terror groups use modern communication technology in myriad ways, from fundraising, radicalization, and recruitment, to issuing threats, inciting violence, and planning attacks. The rapid adoption of state-of-the-art communication tools—with an emphasis on Internet-based applications—has been critical to the organization, expansion, and success of terrorist networks. The early 2000s saw a boom in new media applications that enabled terrorists to communicate undetected across borders more swiftly and effectively. The Internet essentially became another extremist battlefield. Encrypted software became a popular modus operandi for jihadists, and many groups established media departments and online recruitment magazines such as al-Qaeda's *Inspire* and ISIS's *Rumiyah*.

With the emergence of ISIS and its declaration of a caliphate, ISIS leaders have turned to the Internet for radicalization and recruitment. Recruiters utilize social media outlets to "field questions about joining" the group, a process which resembles an "online version of [a] religious seminar." They use social media platforms such as Twitter, Facebook, YouTube, Tumblr, and Ask.fm. For example, analysts estimate that at least 45,000 pro-ISIS accounts were active on Twitter between September-November 2014.[1]

After years of denial and inaction, we are finally seeing some progress by technology companies to tackle extremist activity and terror-related content on the Internet. Most major technology platforms have finally acknowledged the problem, but their promises[2] to do more to combat terrorism online lack in substance. We at the Counter Extremism Project still find terrorist videos on Facebook, YouTube, Google+, and other platforms, and encounter significant gaps in how some companies are dealing with some of the worst content on their platforms.

[1] https://www.counterextremism.com/content/digital-developments-extremists-use-modern-communication-tools
[2] https://www.blog.google/topics/google-europe/four-steps-were-taking-today-fight-online-terror/; https://newsroom.fb.com/news/2018/04/keeping-terrorists-off-facebook/

The Challenge

To illustrate these gaps and the scale of the challenge, Dr. Hany Farid, Senior Advisor to the Counter Extremism Project and Albert Bradley 1915 Third Century Professor and Chair of Computer Science at Dartmouth College, presented research to the European Parliament's Special Committee on Terrorism (TERR) on April 24, 2018. He explained how over a six-week period between March 8 and April 18, 2018, CEP used its own robust hashing technology, eGLYPH, and YouTube's own API to better understand how ISIS content is being uploaded to YouTube, how long it is staying online, and how many views these videos are generating. CEP searched for the presence of a relatively small set of just 256 previously identified ISIS-generated terror videos. Over the 6-week period, using a narrow scope of research parameters, here is what we learned:

1. No less than 942 ISIS videos were uploaded to YouTube.
2. These 942 videos garnered a total of 134,644 views of this small set of terror content.
3. Although 74% of the videos remained on YouTube for less than two hours, within this time window, these videos garnered an average of 12 views each, with a maximum view count of 252.
4. For videos that were available for more than two hours, the average number of views over a 48-hour window was 515, with a maximum view count of 9,589 (we stopped tracking views after a 48-hour window).
5. These 942 videos were uploaded by 157 different YouTube accounts, some of which uploaded as many as 70 videos before the channel was either removed by YouTube administrators or deleted by the user.
6. Approximately 91% of the videos that we found were uploaded more than once and stayed online at least long enough for us to find it.

A few conclusions can be made from this analysis. Despite claims by YouTube that they are using automatic tools to find and remove terror-related content, the same content is repeatedly uploaded, sometimes from the same account. This content is staying online anywhere from hours to days and is being viewed hundreds to thousands of times. Once removed, the same content is then simply re-uploaded, meaning that it is effectively available all the time. The bottom line is that too much terror-related content is finding its way online, it is staying online for too long, and it continues to reappear even if it is eventually taken down. While we have seen some progress as compared to a few years ago, there are still significant gaps in the development and deployment of technology to quickly and accurately find and remove terror-related content.

The eGLYPH Technology

To help tackle this problem, CEP along with Dr. Farid developed eGLYPH—a technological solution that can greatly reduce the ability of extremists and terrorists to spread their content and weaponise online platforms. The technology is capable of detecting known extremist images, video, and audio files through "robust hashing" technology, which was originally deployed to identify and flag images of child pornography online (PhotoDNA), by extracting a distinct digital signature from an image and comparing it against all other images encountered online. eGLYPH expands on this existing technology and is able to analyse video and audio content quickly and accurately, making it particularly impactful in combatting the proliferation of extremist propaganda.

eGLYPH can be deployed in two ways:

1. Internet and social media companies can deploy eGLYPH on their platforms to detect content and flag for removal at the point of upload.

2. eGLYPH can be attached to a web crawler to actively scrape the Internet for content. As content is detected, takedown notices can be submitted to companies to request immediate removal of the flagged content.

The most efficient detection works when a person initially identifies images, videos, or audio recordings for removal and then eGLYPH extracts a distinct digital signature, or "hash," from this database of content. This signature is then used to find duplicate uploads. Matches are reported to the company, and if the content violates the terms of service, it is removed. Companies already work to take down content that is violent, horrific, and violates their terms of service. CEP's eGLYPH serves to streamline and accelerate the flagging and removal process in the online extremism space.

CEP's Recommendations

Online platforms are not doing enough to tackle extremist content online despite the technology industry's efforts to convince policymakers, corporate advertisers, the media, and civil society otherwise. Tech companies, especially multi-billion-dollar firms like Google/YouTube and Facebook, should take responsibility as industry leaders and dedicate the requisite resources and capabilities to removing this dangerous material. Online platforms must do more to improve existing technologies and develop and deploy new technology to contend with an ever-changing Internet landscape and continued presence of extremist content.

Social media companies have begun to act but much more can be done. We recommend that:

1. It is of utter importance to use automated technology along with the necessary human verification capabilities when identifying and tackling illegal content online. Hashing technology like eGLYPH, when coupled with human reviewers, is most effective and reliable in taking down known extremist content. Human researchers and content moderators must be included in the decision-making process and be on the lookout for emerging trends. Technology companies should ensure that these departments are fully staffed, and appropriately prepare, train, and educate its human reviewers.

2. Technology firms must be more transparent regarding its hashing efforts. Google/YouTube, Facebook, and Twitter should fully explain how it is implementing hashing technology, specifically if they are deploying at the point of upload. These companies should also provide a detailed explanation of how each contributes to and participates in the so-called "hashing coalition" announced in December 2016. Each company should state how much content they have contributed to this shared database, and whether there is an agreement that all content in the database be removed across industry platforms and websites that are members of the hashing coalition and the Global Internet Forum to Counter Terrorism (GIFCT). Google/YouTube, Facebook, and Twitter should also state how much content has been removed from their platform as a result of the database, and explain how the database is updated. An objective of the GIFCT is to share knowledge, information, and best practices. As members, Google/YouTube, Facebook, and Twitter should aim to set industry standards on hashing practices. A

mandate for all GIFCT members to hash and remove content produced by groups and individuals as sanction-designated by the United States, European Union, and United Nations, as well as material that glorifies or incites violence, would rationalize content removal practices and dramatically reduce the amount of terrorist content online.

3. Platforms must be proactive in content monitoring. Many social media sites primarily review and remove content that has been reported to them. Instead, given the vast financial resources of major technology companies, each should spearhead internal efforts to find content and remove it without relying on the public to police the platform for them. Revenues accrue exclusively to the technology company, and users should not be solely responsible for reporting problems of a platform to the company.

4. While removing content quickly from Internet and social media sites is clearly an important component of any effort to restrict the dissemination of terrorist propaganda, "time online" should not be the only metric used to gauge tech's progress in combatting terrorist propaganda. CEP has found that in many cases, ISIS videos removed within two hours still received dozens and, in some cases, hundreds of views.

Policymakers, advertisers, and the public must continue to pressure technology companies to take more responsibility for the direct and measurable harm coming from the abuses on their platforms. It is clear that the technology industry takes action in response to highly publicized discoveries of extremist content on their platforms as well as threats of regulation or loss of advertising revenue. If these technology companies do not respond more effectively, then policymakers should consider fines (as the Germans have) and advertisers should consider withholding advertising dollars (as Unilever has threatened to do).