Royal United Services Institute
for Defence and Security Studies

Swansea University
Prifysgol Abertawe

Hillary Rodham Clinton
School of Law | Ysgol y Gyfraith

Global Research Network on Terrorism and Technology: Paper No. 3

Shedding Light on Terrorist and Extremist Content Removal

Isabelle van der Vegt, Paul Gill, Stuart Macdonald and Bennett Kleinberg

Social media and tech companies face the challenge of identifying and removing terrorist and extremist content from their platforms. This paper presents the findings of a series of interviews with Global Internet Forum to Counter Terrorism (GIFCT) partner companies and law enforcement Internet Referral Units (IRUs). It offers a unique view on current practices and challenges regarding content removal, focusing particularly on human-based and automated approaches and the integration of the two.

Summary of Recommendations

- The major technology platforms should begin developing hybrid systems (if they have not done so already) for their own and others' use.
- Technology companies should actively promote the availability, effectiveness and transparency of appeal procedures. This includes providing to users as full an explanation as possible of why their content was removed.
- The GIFCT shared database of hashes – which focuses on 'the most extreme and egregious terrorist images and videos'[1] – should be expanded to include more borderline content.

Introduction

In the digital age, counterterrorism involves social media and technology companies tackling the spread of terrorist and extremist content on their platforms. This engenders lengthy debates around the importance of effective and rapid responses to this issue within media, government and the general public. This is particularly the case in the aftermath of major terrorist attacks such as the Christchurch Mosque shootings in New Zealand. The authors' assumption is that the key to effectiveness lies in the ability to improve hybrid human machine decision-making. To do so, however, requires an understanding of exactly how judgements to remove or not remove content are reached in both systems (human and machine).

This paper offers insights from in-depth interviews with representatives from law enforcement and small- and large-scale tech companies on the subject of terrorist and extremist content removal. The interviews aimed to:

1. Understand how the removal of material for the promotion of terrorism is handled.
2. Examine which criteria are deemed essential to flag content as 'terrorist'.

[1] *Facebook Newsroom*, 'Partnering to Help Curb Spread of Online Terrorist Content', 5 December 2016, <https://newsroom.fb.com/news/2016/12/partnering-to-help-curb-spread-of-online-terrorist-content/>, accessed 27 June 2019.

3. Discern the properties of online content that are used to make the remove/not remove decision.

In effect, the interviews allow the authors to describe current practices and challenges in the field, particularly focusing on human-based and automated approaches to content removal and the integration thereof. Six themes emerged from the interviews and each are elaborated upon in detail in this paper.

1. The two major approaches to content removal.
2. Different types of violating content.
3. The importance of policy guidelines.
4. The importance of human decision-making.
5. Differing approaches to measuring accuracy.
6. The problem of 'grey-zone' content.

Thereafter, two discussion points are presented that arose from the identified themes and the paper concludes by condensing these findings into three policy recommendations.

Method

GIFCT partner companies and law enforcement based IRUs were approached to participate in semi-structured interviews.[2] Questions centred around the removal procedures of terrorist and extremist content, specifically the division between manual and automated decisions for removal. Interview questions and an information sheet were sent to interview participants.[3] Interviewees included company (counterterrorism) policymakers, public policy strategists and IRU managers, with seven interviews conducted with nine interviewees. Each interview lasted between 20 and 60 minutes, was digitally recorded and fully transcribed, and the content was examined qualitatively for recurrent themes which are outlined in the introduction. These themes are discussed in depth in the following sections.

The Two Major Approaches to Content Removal

The content-removal process is largely driven by measures focusing on either content or behaviour.

Content-based decisions focus on linguistic characteristics, word use, images, and URLs. Behaviour-based decisions look at indicators of the account (for

[2] Interviews were conducted on the understanding that companies and interviewees would not be named, and information would not be attributed to specific interviewees.

[3] Open Science Framework, 'GIFCT: Terrorist Content Removal', 23 March 2019, <https://osf.io/7jtd2/>, accessed 27 June 2019.

example, how long ago it was opened, how often it posts messages) or message behaviour (for example, including trending or unrelated hashtags) independent of the content.

Various social-media companies have hired human content moderators, while the larger companies have hired up to several thousand, who manually decide whether specific content violates company policies. There is a triaging process in place that combines human expertise with automated judgements: an automated terrorist content detection system can flag suspect messages or accounts for humans to review. Equally, human judgements are fed back into automated systems. Within law enforcement, teams of specialists (for example, on specific languages and cultures of interest) are also tasked with judging pieces of content and referring suspect cases to social-media companies.

Content-based decisions rely heavily on human involvement. Scaling an approach that focuses on content cues is therefore difficult, particularly for smaller companies. However, decisions to remove content are sometimes also made by other means. One of the interviewed parties likened the removal process to spam filtering, where behavioural cues determine whether a piece of content or an account is removed. Behavioural cues may include abnormal posting volume (for example, several posts per minute), or tagging a single post with various trending hashtags to gain attention. Such cues can be picked up with relative ease by automated systems, and often will not require any human intervention. Indeed, the same interviewee raised the point that terrorist and extremist entities exhibit specific online behavioural patterns (for example, rapid content release) in an attempt to reach a large audience before any systems detect that terrorist propaganda or other terrorist content has been posted. This type of behaviour was referred to as a ‘raid model’, where an individual opens the account knowing that it will be taken down soon, and thus must resort to posting a large amount of content in a short time. With a behavioural, content-agnostic approach, an account may even be flagged before any content has been posted, because its behaviour (such as the accounts it follows) may already resemble spam behaviour and suggest a terrorist or extremist entity.

Different Types of Violating Content

In 2015, Donald Holbrook demonstrated the broad spectrum of terrorist group content.[4] At the lower end of the scale, ‘moderate’ content may contain no endorsement of violence or hatred towards identified communities. ‘Fringe’ content may demonstrate anger and hostility towards a given group of people without the added assumption that these people are somehow ‘subhuman’ and legitimate targets of violence. In incremental steps, ‘extremist’ content

[4] Donald Holbrook, ‘Designing and Applying an “Extremist Media Index”’, *Perspectives on Terrorism* (Vol. 9, No. 5, 2015), pp. 57–68.

might: glorify general collective violence; glorify violence against specific groups; provide facilitation, scope or direction of specific acts of violence against specific targets; and provide specific instruction on how to commit types of violence. Each provides a challenge for moderation and the threshold required for an intervention may differ across companies and jurisdictional boundaries. Indeed, some IRUs are mostly concerned with removing propaganda that bears the ‘brand’ of a terrorist organisation. In fact, some of this content is far from being considered violent. Interviewees also raised the importance of monitoring and removing so-called ‘utopian’ content, that is, texts, images and videos that praise or glorify extremist lifestyles (for example, showing unrealistically peaceful scenes from bombarded regions). These forms of propaganda may be as dangerous as graphically violent pieces because they may similarly – or possibly more strongly – mobilise people into action, as stated by one interviewee. The violent and graphic content often associated with terrorist and extremist materials (for example, beheadings) may be easiest to judge since there is no question about whether it should be taken down. In contrast, veiled propaganda and support is more challenging.

Another example of challenging content relates to the distinction between terrorists and political actors. One interviewee mentioned an example where a group carried out attacks against infrastructure without the intent to harm civilians (by releasing warnings). After deliberation, the company decided not to remove that content because it was judged as not meeting their criteria for terrorist content. However, since different definitions of terrorism exist, other companies may well have opted for removal in such a case, regardless of the intent to not harm civilians.

To summarise, violent images and videos are just one part of a range of materials that, in the judgement of social media companies and law enforcement organisations, need to be removed. It is equally, if not more, important to tackle content that promotes a violent extremist group’s utopian ideal or glorifies extremism. This also holds for cases where the (violent) intent of a piece of content is purposefully hidden or coded.

The Importance of Policy Guidelines

Many law enforcement agencies established their own content moderator teams – often designated as IRUs. They flag to social media companies pieces of content they find particularly worrying. IRUs do not make the final decision on whether content is removed, nor do they enforce their decision. Upon receiving such a referral, most platforms will check whether the specific instance violates its terms and conditions, leaving open the possibility of disagreeing with an IRU’s referral. Indeed, all GIFCT partners have their own specific terms of service that bar terrorist entities from a presence on their platform, and similarly will not allow individual users to praise, support or represent such organisations.

Implementing such policy guidelines on smaller platforms is equally, if not more, important. In the past, there have been instances of smaller platforms being confronted with an influx of terrorist users once the platform had been identified as an easy or convenient target for spreading content. Content is only removed from platforms if it violates policy guidelines, even if it has been flagged by law enforcement agencies. Therefore, it seems important that technology companies have comprehensive terms of service in place to counter a wide range of terrorist and extremist content.

The Importance of Human Decision-Making

All interviewees reported that a human element is integral to their decision-making process. In fact, most smaller platforms reported not using any form of automation. Human moderators are often tasked with examining 'grey zone' content, where materials may be judged by several people before a decision is made, or where a team of moderators and/or policymakers will discuss what type of action should be taken. Human moderators are also needed to judge the context surrounding pieces of terrorist and extremist content. Some pieces of content may, for example, appear in a journalistic context, where a news organisation is factually reporting on a piece of content, condemning it, or sharing (parts of) it to undermine it.

Other contextual nuances, such as cultural and political specificity, coded language, humour, satire, and irony are also better judged by humans, often with specialist knowledge on the topic, language, cultural or political situation. This, for example, includes understanding new emerging groups, but also distinguishing between content relating to a military conflict versus a terrorist attack.

A common assertion made by interviewees about human content moderators was that they are domain experts (for example, with professional expertise in counterterrorism), linguists (for example, those who understand highly localised language), and area experts (for example, those who are aware of the groups operating in a small geographical region and have on-the-ground knowledge). Another reason interviewees identified for relying on human expertise is the ability to identify 'adversarial shifts', where groups change their modus operandi or the targeted platform. Such shifts in the content-spreading strategies of terrorist groups prove problematic. As mentioned by one of the interviewees, in a smaller-scale operation automated systems take longer to develop due to limited resources. In addition, due to adversarial shifts such systems may lose their relevance once implemented and once personnel are trained in their use.

It is important to note that automated content-removal efforts were predominantly used by the larger social media platforms and not by IRUs or small platforms. Smaller platforms resorted to manual approaches to collect suspect content and remove it – often as a judgement call or after flagging from other platform users. A possible reason is the number of resources needed to establish engineering teams that build automated content-removal systems. When automated systems were in place, one of the interviewed parties emphasised the importance of an appeal process. The appeals process, in general, includes the option of a user raising concerns about the blocking of their account or the removal of (some of) their posted content. When removal decisions are challenged, they are reviewed by human experts. One desirable effect that one of the companies hopes to achieve with the appeal process is handling false-positive decisions in a less restrictive manner.

All in all, human decision-making remains integral to every platform whose representatives were interviewed. Multiple interviewees stated that human expertise trumps automated systems in instances where expert knowledge is necessary. In some cases, human moderation was the sole system in place. Where automated systems are used, mostly in large-scale operations, the importance of an adequate appeal process was emphasised, so that potential false positives can be re-reviewed.

Differing Approaches to Measuring Accuracy

A recurrent theme in the public discourse about content removal is that of removal accuracy. Indeed, an interviewee from one of the large-scale companies that makes use of automated systems stressed the importance of reaching removal accuracy rates ‘as close to 100%’ as possible. However, there seems to be no consensus in the field as to how this accuracy can most effectively be measured. One option that was raised by two different interviewees is examining the appeals rate, where the number of successful appeals (false positives) may reflect the quality of the removal system. Another route mentioned by an interviewee involves internally reviewing content, where human moderators from different levels of expertise re-label a random sample of content that was taken down. Furthermore, it was also mentioned by one interviewed party that automated systems can be used to carry out more fine-grained photo- and video-matching, which can produce a figure of how accurate the matching process is. For IRUs, accuracy rates can be measured in terms of the rate of flagged content that is actually taken down by social-media companies. Regarding this, one IRU manager estimated that 99% of the content that is referred to social-media companies by the unit is indeed removed. All in all, no standardised measure of estimating accuracy in terrorist and extremist content removal emerged from the interviews.

The Problem of ‘Grey-Zone’ Content

Many interviewees raised the problem of the judgement ‘grey zone’, which refers to those cases where content is not obviously terrorist content but not clearly innocent either. For example, one interviewee described these cases as ones that would be considered ‘uncomfortable’ or ‘bad speech that is not bad enough to meet the takedown’. For these cases, in particular, the low base rate of actual terrorist content presents a dilemma: with a high chance for false positive decisions, the removal of grey-zone content can be interpreted as over-censorship. Conversely, if some consider grey-zone content to be terrorist content, failure to remove it might result in public outcry and alleged violations of the 'Nine Steps' that leading social media companies pledged to implement in response to the Christchurch Call.[5] Several interviewees reported that they err on the side of freedom of speech rather than over-censorship. That effectively means that the problem is currently solved by not removing doubtful cases. While that decision does lower the false positive rate, it may also be perceived as an increase in false negatives (in other words, terrorist content that is not removed).

Discussion

The preceding sections outline current practices in terrorist and extremist content removal, as well as common challenges faced within the field. The issues identified in this research give rise to two discussion points, which are condensed into policy recommendations in the conclusion.

First, it is worthwhile considering the argument that highly accurate automated tools will solve the problem. Indeed, an interviewee from a large-scale company stated that only highly accurate (close to 100% accuracy rates) systems are put into place. Even if such accuracy rates are a possibility, problems with false positives and negatives will persist. A popular misperception is that as soon as the content-removal systems become more accurate, human expertise becomes obsolete and the problem will disappear. Consider an example of a company that took down 1 million pieces of what it deemed to be terrorist content. Suppose automated systems identify terrorist content with 95% accuracy and identify unproblematic content with the same accuracy. Intuitively, one would concur that if a piece of content is removed, the chances that this was indeed terrorist content are 95%. But this is far from correct because of a statistical catch. Since most content is non-terrorist (for the sake of simplicity, assume 1% of content is terrorist), the relatively small error rate of 5% still amounts to an enormous absolute number.

Under the assumption that a large social-media company deals with 100 million pieces of content in a given period, three important calculations can be made: 1) it is known that there are 1 million pieces of actual terrorist content in the sample (1%); 2) with a 95% accuracy rate, 950,000 pieces of actual terrorist content will be accurately identified (true positives), and 50,000 pieces will not be recognised (false negatives); and 3) of the 99 million normal content pieces, 4.95 million (5%) are still falsely identified as terrorist content (false positives), and 94.05 million will be correctly identified as

[5] See Amazon, Google, Microsoft, Facebook and Twitter, 'Joint Statement in Support of Christchurch Call', <https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2019/05/Christchurch-Call-and-Nine-Steps.pdf>, accessed 1 July 2019.

non-terrorist (true negatives).[6] In short, 5.9 million pieces in total will be flagged as terrorist content (false positives + true positives = 4.95 million + 950,000). This means that if a piece of content was removed, the chances that it was in fact terrorist content are merely 16.10% (true positives/false positives + true positives = 950,000/5.9 million). This base rate fallacy renders the task of content removal highly challenging: even with outstanding accuracy rates, companies will by the very nature of the problem end up removing mainly innocent content. Added to that problem is the difficulty of measuring accuracy in the first place, where different companies have different understandings of true and false positives and negatives.

Second, when it comes to grey-zone content, the solution might not lie solely in automated approaches because of the nuances and context that create the grey-zone problem in the first place. Rather, it might simply be that this issue cannot currently be addressed. This is partly because its difficulty might have been underestimated, and partly because the ability of tech companies to solve it might have been overestimated. However, this is mostly because the problem by definition is not solvable to the satisfaction of all stakeholders. Government, law enforcement, tech companies and the general public may all have different interests and standards when it comes to extremist content removal. Moreover, different stakeholders abide by different definitions of terrorism. Similarly, law enforcement bodies may use different definitions depending on the law of the state in which they are located.

Instead, what is needed is research into collaborative decision-making processes that more effectively combine multiple layers of automated content moderation (for example, content-based approaches and behavioural approaches) and human expertise. A cascaded approach could be implemented, whereby each layer narrows down the number of content pieces that need to be further assessed for containing terrorist content. Ideally, there remains in the last layer of the automated cascade only a small proportion of the content (for example, those in the 'grey zone') to be reviewed by human moderators. However, the indicators for each layer of the assessment procedure need to be independent of (in other words, uncorrelated with) each other, in order to not propagate false positive and false negative rates. In short, if decision-making strategies that signal terrorist content independently from one another were to be found (for example, two cues that signal terrorist content in different ways), collaborative cascaded decision-making approaches might offer a path towards better online regulation.

[6] Note that the accuracy rate of 95% can then be re-calculated as follows: (950,000 + 94.05 million)/(95,000 + 94.05 million + 4.95 million + 50,000) = (true positives + true negatives) / (true positives + true negatives + false positives + false negatives) = 0.95.

Recommendations

This paper offers a view on various approaches to extremist and terrorist online content removal. Through interviews with representatives of leading technology companies and law enforcement agencies, the authors have shed light on current practices in this field. More importantly, the interviews have enabled the authors to identify avenues for future research and suggestions for industry improvement. From an analysis of the interviews, the authors propose three policy recommendations:

1. The problems of contextual ambiguity and grey-zone decisions remain issues that need to be addressed through effective hybrid systems, integrating both human and automated content moderation, where collaborative, cascaded decision-making systems combine independent indicators. **The major technology platforms should begin developing hybrid systems (if they have not done so already) for their own and others' use.**
2. Due to the base rate fallacy, even highly accurate systems are faced with the problem of false positives. Therefore, straightforward appeal processes are a necessary safeguard. Moreover, the opportunity to appeal should be a meaningful one. **Technology companies should actively promote the availability, effectiveness and transparency of appeal procedures. This includes providing to users as full an explanation as possible of why their content was removed.**
3. Specialist knowledge is required to understand the nuances of much extremist (and borderline) content. Such knowledge may not always be in place, especially within smaller companies, meaning false positives and false negatives are more likely. Collaboration between large- and small-scale tech companies is therefore essential. Knowledge-sharing will improve the ability of companies with fewer resources to effectively counter terrorist and extremist presence on their platforms. To this end, **the GIFCT shared database of hashes – which focuses on 'the most extreme and egregious terrorist images and videos'[7] – should be expanded to include this more borderline content.**

Isabelle van der Vegt is a PhD candidate in the Department of Security and Crime Science at University College London.

Paul Gill is a Professor in the Department of Security and Crime Science at Univeristy College London.

Stuart Macdonald is a Professor of Law in the School of Law at Swansea University.

Bennett Kleinberg is a Lecturer in the Department of Security and Crime Science at University College London.

[7] *Facebook Newsroom*, 'Partnering to Help Curb Spread of Online Terrorist Content'.

About RUSI

The Royal United Services Institute (RUSI) is the world's oldest and the UK's leading defence and security think tank. Its mission is to inform, influence and enhance public debate on a safer and more stable world. RUSI is a research-led institute, producing independent, practical and innovative analysis to address today's complex challenges.

Since its foundation in 1831, RUSI has relied on its members to support its activities. Together with revenue from research, publications and conferences, RUSI has sustained its political independence for 188 years.

About The Global Research Network on Terrorism and Technology

The Global Research Network on Terrorism and Technology is a consortium of academic institutions and think tanks that conducts research and shares views on online terrorist content; recruiting tactics terrorists use online; the ethics and laws surrounding terrorist content moderation; public-private partnerships to address the issue; and the resources tech companies need to adequately and responsibly remove terrorist content from their platforms.

Each publication is part of a series of papers released by the network on terrorism and technology. The research conducted by this network will seek to better understand radicalisation, recruitment and the myriad of ways terrorist entities use the digital space.

The network is led by the Royal United Services Institute (RUSI) in the UK and brings together partners from around the world, including the Brookings Institution (US), the International Centre for Counter-Terrorism (Netherlands), Swansea University (UK), the Observer Research Foundation (India), the International Institute for Counter-Terrorism (Israel), and the Institute for Policy Analysis of Conflict (Indonesia).

The research network is supported by the Global Internet Forum to Counter Terrorism (GIFCT). For more information about GIFCT, please visit https://gifct.org/.

The views expressed in this publication are those of the authors, and do not necessarily reflect the views of RUSI or any other institution.

Published in 2019 by the Royal United Services Institute for Defence and Security Studies. RUSI is a registered charity (No. 210639).

Royal United Services Institute
for Defence and Security Studies
Whitehall
London SW1A 2ET
United Kingdom
+44 (0)20 7747 2600
www.rusi.org

RUSI is a registered charity (No. 210639)