Tackling terrorist content online – Propaganda and content moderation

Whitepaper

www.notiones.eu office@notiones.eu @NOTIONES_EU NOTIONES Project

tate.techagainstterrorism.org/ Tech Against Terrorism Europe

Authors

Professor Stuart Macdonald,
Director, Cyber Threats Research Centre (CYTREC),
Swansea University & TATE project partner
s.macdonald@swansea.ac.uk

Andrew Staniforth,
Director of Innovation, SAHER (Europe),
TATE Project Coordinator & NOTIONES project partner
andy@saher-eu.com

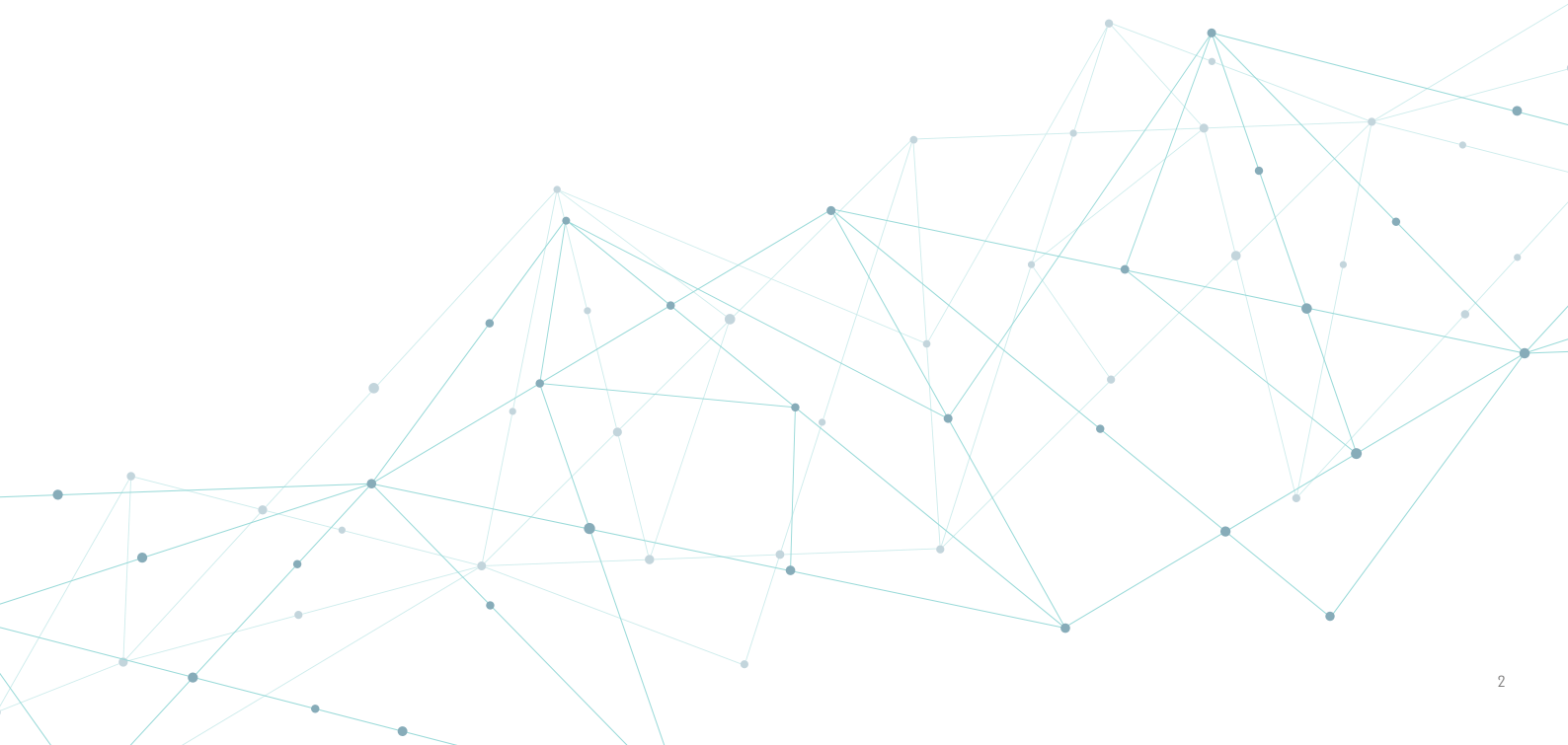

Table of Contents

Project Introduction: NOTIONES

Novel technologies have presented practitioners with new opportunities to improve the intelligence process, but have also created new challenges and threats. Consequently, the timely identification of emerging technologies and analysis of their potential impact, not only on the intelligence community but also on terrorist or criminal organisations, is crucial.

However, time constraints can prevent intelligence practitioners from being updated on the most recent technologies.

In order to address this challenge NOTIONES will establish a network, connecting researchers and industries with the intelligence community. This network will facilitate exchange on new and emerging technologies but also equip solution providers with insights on the corresponding needs and requirements of practitioners. The so gained findings will be disseminated in periodic reports containing technologic roadmaps and recommendations for future research projects and development activities.

The consortium of NOTIONES includes, among its 29 partners, practitioners from military, civil, financial, judiciary, local, national and international security and intelligence services, coming from 9 EU Members States and 6 Associated Countries. These practitioners, together with the other consortium members, grant a complete coverage of the 4 EU main areas: West Europe (Portugal, Spain, UK, France, Italy, Germany, Austria), North Europe (Finland, Denmark, Sweden, Estonia, Latvia), Mittel Europe (Poland, Slovakia, Ukraine), Middle East (Israel, Turkey, Georgia, Bulgaria, Greece, North Macedonia) for a total of 21 countries, including 12 SMEs with diverse and complementary competences.

Project Objectives

GATHER the needs of intelligence and security practitioners related to contemporary intelligence processes and technologies;

PROMOTE interaction of technology providers and academy with intelligence and security practitioners;

IDENTIFY novel technologies of relevance for practitioners through research monitoring;

PUBLISH a periodic report, summarising key findings in order to orientate future research and development;

ENSURE the commitment and involvement of new organisations in the pan-European NOTIONES network.

Project Introduction: NOTIONES

Project Facts:

Duration: 60 Months **Reference: 101021853**

Programme: Horizon 2020 SU-GM01-2020 Coordination and Support Action

Coordinator: FUNDACION TECNALIA RESERACH & INNOVATION (Spain)

Scientific Technical Coordinator: ZANASI ALESSANDRO SRL (Italy)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 101021853.

Coordinator

Scientific Technical Coordinator

Project Security Officer

Academic | Think-Tanks | Research

Technology Providers

Practitioners

Project Introduction: TATE

Tech Against Terrorism Europe (TATE) will support smaller hosting services providers (HSPs) in preventing terrorist actors from disseminating terrorist content as defined in the EU's terrorist content online (TCO) regulation and in Directive (EU) 2017/541. Combining unique industry-leading expertise from private sector organisations and leading academic institutions actively engaged in tackling TCO, the consortium of partners will ensure TATE delivers the long-term impacts of large-scale disruption of TCO on priority HSPs, providing a sustainable foundation for practical support mechanisms for smaller HSPs in countering terrorist content online.

The mission of TATE will be achieved by increasing awareness of the TCO Regulation and requirements among small HSPs through the creation of a series of unique interactive learning materials. This will be supported by the introduction of a bespoke TCO capacity-building programme for HSPs, taking priority HSPs through the capacity building programme, scaling existing technical solutions to benefit all smaller HSPs in scope for the TCO regulation.

Project Objectives

INCREASE awareness about the TCO Regulation and requirements among small HSPs by creating a series of written and interactive learning materials;

AMPLIFY the understanding of HSPs of the legality and taxonomy of terrorist-related content to ensure the important preservation of removed content for future LEA analysis, assessment and investigation;

INCREASE the number of small HSPs that implement the TCO Regulation effectively including the removal of terrorist content within 1 hour;

ESTABLISH contacts between small HSPs to exchange best practices among each other via the organisation of workshops and allowing for communication via existing infrastructure;

INCREASE the volume of online terrorist content removed by small HSPs and enhance their communication with competent authorities.

Project Introduction: TATE

Project Facts:

Duration: 24 months **Reference:** 101080101

Programme: Internal Security Fund Terrorist Content Online (ISF-2021-AG-TCO-101080101)

Coordinator: SAHER (Europe) OU

This project has received funding from the European Union's Internal Security Fund 2021 Terrorist Content Online call under Grant Agreement No 101080101.

1. Introduction

In his recently published report *The Terrorism Acts in 2021*, the UK's Independent Reviewer of Terrorism Legislation stated that 'most terrorism arrestees are profoundly engaged in expressing and consuming violent and hateful material online, and that online encouragement can be troublingly effective at promoting violence in others'.[1] This has also been the experience of counterterrorism police.[2] A recent study of individuals convicted of extremism offences in the UK provides empirical support for this view, concluding that the internet is playing an increasingly prominent role in radicalisation processes and that radicalisation now takes place primarily online.[3]

In the light of these findings, the focus of this whitepaper is the response of the tech industry to online terrorist and violent extremist content (TVEC), which serves to inform the ongoing research and innovation activities of EU funded projects TATE (Tech Against Terrorism Europe) and NOTIONES (iNteracting netwOrk of inTelligence and securIty practitiOners with iNdustry and acadEmia actorS), being of direct interest and operational value to the multidisciplinary stakeholders operating across the counterterrorism and intelligence landscape.

The whitepaper has three parts. The first part provides some contextual background, describing the diverse range of online services utilised by terrorists and extremists and the process by which propaganda is disseminated online. The second part details industry responses. As well as referrals from users and law enforcement, it describes the use of AI for proactive detection and collaborative, cross-platform initiatives. The third part describes four issues for discussion: transparency; definitional clarity; the impact on those targeted; and, the use of online data for predictive purposes.

[1] Jonathan Hall, The Terrorism Acts in 2021: Report of the Independent Reviewer of Terrorism Legislation on the Operation of the Terrorism Acts 2000 and 2006, and the Terrorism Prevention and Investigation Measures Act 2011 (His Majesty's Stationery Office, 2023), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1140911/E02876111_Terrorism_Acts_in_2021_Accessible.pdf, accessed March 18, 2023, 160.

[2] Stuart Macdonald and Andrew Staniforth, Tackling Online Terrorist Content Together: Counterterrorism Law Enforcement and Tech Company Cooperation, (London: Global Network on Extremism and Technology, 2023), https://gnet-research.org/wp-content/uploads/2023/01/31-Tackling-Online-Terrorist-Content-Together_web.pdf, accessed March 19, 2023.

[3] Jonathan Keynon, Jens Binder and Christopher Baker-Beall, The internet and radicalisation pathways: technological advances, relevance of mental health and role of attackers (HM Prison & Probation Service, 2022), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1121985/internet-radicalisation-report.pdf, accessed March 19, 2023.

2. Context

A vast amount of terrorist content is posted to the biggest social media platforms every day. In 2021 alone, Facebook removed more than 34 million items of terrorist propaganda, YouTube removed 513,908 videos for the promotion of violence and violent extremism, and Twitter suspended 78,668 accounts for the promotion of terrorism.[4] Given these figures, it is important that there continues to be scrutiny of the efforts of the biggest social media companies to tackle TVEC on their platforms. At the same time, it is also necessary to widen the lens. This section begins by highlighting the variety of different service types that are utilised in terrorist and extremist online ecosystems, pointing in particular to the need to develop a strategy for tackling terrorist operated websites. It then describes the propaganda dissemination strategy employed by Islamic State (IS), in order to highlight the exploitation of (often small or micro) file-sharing platforms.

a. The online ecosystem

It is important to recognise the variety of different online services that are exploited by terrorists and extremists. A study of the ecosystems of two European far-right online networks identified eleven different types of service. As well as social networking, these service types included websites, video sharing, follower tracking, URL shortening, social media marketing/posting/sharing, online petitioning, internet archiving and video streaming.[5] Other studies have yielded similar results.[6] This diversity is also illustrated by the list of members of the Global Internet Forum to Counter Terrorism (discussed further in section 3c). As well as the founding members Facebook, Twitter, YouTube and Microsoft, other members include such companies as WordPress, Amazon, MailChimp, AirBnB, GIPHY and the file-sharing site JustPaste.it.

Recent analyses have urged the importance of combatting terrorist operated websites.[7] While terrorists and extremists rely less on websites than they once did, websites still play an important role in the online ecosystem and 'could re-emerge more strongly with accelerated disruption of extremist and terrorist content and accounts by social media platforms and adjacent services unless providers further down "the tech stack" take more concerted action'.[8] There are several reasons why terrorists might find a website appealing.[9] Websites can function as archives of content. Unlike social media, website content is often indexed by search engines. And users retain greater control over the content of their websites. In early 2022, Tech Against Terrorism reported that since the start of 2020 it had identified a total of 198 websites operated by terrorists or violent extremists.[10] Further analysis of a sample of 33 of these websites found that in total they had 1.54 million monthly visitors. 91% of the sites displayed propaganda and 57% included a contact address form. Six months later, Tech Against Terrorism had identified 14 more sites. It stated that 'this issue is largely absent from government-led policy discussions on disrupting terrorist use of the internet. As a result, there is no common global mitigation strategy'.[11]

[4] 'Community Standards Enforcement Report – Dangerous Organizations: Terrorism and Organized Hate', https://transparency.fb.com/data/community-standards-enforcement/dangerous-organizations/facebook/#content-actioned, accessed February 11, 2023; 'YouTube Community Guidelines Enforcement', https://transparencyreport.google.com/youtube-policy/removals, accessed February 11, 2023; 'Rules Enforcement', https://transparency.twitter.com/en/reports/rules-enforcement.html, accessed February 11, 2023.

[5] Stuart Macdonald et al, ***The European Far-right Online: An Exploratory Twitter Outlink Analysis of German & French Far-Right Online Ecosystems***, (Washington, DC: Resolve Network, 2022), https://doi.org/10.37805/remve2022.2.

[6] ***Transparency Report: Terrorist Content Analytics Platform, Year One: 1 December 2020 – 30 November 2021,*** (London: Tech Against Terrorism, 2022), https://www.techagainstterrorism.org/wp-content/uploads/2022/03/Tech-Against-Terrorism-TCAP-Report-March-2022_v6.pdf, accessed March 18, 2023.

[7] Using the term website to refer to a standalone, largely non-interactive, multimedia site.

[8] Maura Conway and Seán Looney, *Back to the Future? Twenty First Century Extremist and Terrorist Websites*, (Luxembourg: European Union, 2021), https://home-affairs.ec.europa.eu/system/files/2022-03/Terrorist%20Operated%20Websites%20Workshop-paper.pdf, accessed March 18, 2023, 3.

[9] ibid.

[10] ***The Threat of Terrorist and Violent Extremist-Operated Websites,*** (London: Tech Against Terrorism, 2022), https://www.techagainstterrorism.org/wp-content/uploads/2022/01/The-Threat-of-Terrorist-and-Violent-Extremist-Operated-Websites-Jan-2022-1.pdf, accessed March 18, 2023. 101 were linked to the far-right; the other 97 were jihadist.

[11] ***Responding to Terrorist Operated Websites, (London: Tech Against Terrorism, 2022)***, https://www.techagainstterrorism.org/wp-content/uploads/2022/07/TAT-TOW-Mitigation-Strategy-July-2022.pdf, accessed March 18, 2023, 1.

2. Context

The mitigation strategy proposed by Tech Against Terrorism seeks to engage four types of web infrastructure: search engines; web hosting providers; domain name system registrars; and, DNS registries.[12] One of the biggest challenges facing any such strategy is that removed websites may reappear, hosted by a different provider or DNS registrar. A further complicating factor is the multi-jurisdictional and cross-sector dimension: 'there are jurisdictional gaps between governments, within governments, and between governments and tech companies as to who should lead, request, and coordinate action'.[13]

b. Propaganda dissemination strategies

IS enjoyed its so-called 'Golden Age' on Twitter in 2013 and 2014.[14] According to one study, in late 2014 there were between 46,000 and 90,000 overt IS supporter accounts on Twitter.[15] These accounts posted an average of 7.3 tweets per day.[16] As enforcement activity increased, and Twitter became a more hostile environment, IS's community-building activities were driven to other platforms, particularly Telegram.[17] Telegram has been found to be used for a variety of purposes by pro-IS users, including instruction, interaction and communication, but by far the most common purpose for which it is used is the distribution of core IS media and other pro-IS materials.[18] Other jihadist and far-right groups have used Telegram in a similar way.[19]

Telegram is a cross-platform messaging app on which users can share an unlimited number of photos, videos and files, of up to 2 gigabytes each.[20] It has over 500 million active users[21] and is popular for its enhanced privacy and encryption.[22] Its features include: secret chats, with end-to-end encryption; a self-destruct timer that permanently deletes secret messages after a set period of time; groups, which are multi-person chats and can have up to 200,000 members; and, of particular relevance, channels, which are a tool for broadcasting messages to large audiences and can have an unlimited number of subscribers.[23] Channels can be public or private. Public channels have a username, so anyone can find them in Telegram's search function and join, whereas to join a private channel a user must be added by the owner or receive an invite link (known as a joinlink).[24]

When a new item of official IS propaganda is produced, it is posted in private Telegram channels.[25]

[12] Ibid.
[13] Ibid, 4.
[14] Maura Conway *et al*, 'Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts,' *Studies in Conflict & Terrorism* 42, no. 1-2 (2019): 150, https://doi.org/10.1080/1057610X.2018.1513984.
[15] JM Berger and Jonathon Morgan, *The ISIS Twitter Census: Defining and describing the population of ISIS supporters on Twitter*, (Washington, DC: Brookings Institution, 2015), https://www.brookings.edu/wp-content/uploads/2016/06/isis_twitter_census_berger_morgan.pdf, accessed March 18, 2023.
[16] ibid.
[17] Nico Prucha, 'IS and the Jihadist Information Highway – Projecting Influence and Religious Identity via Telegram,' *Perspectives on Terrorism* 10, no. 6 (2016): 48–58; Audrey Alexander, *Digital Decay? Tracing Change Over Time Among English-Language Islamic State Sympathizers on Twitter*, (Washington, DC: George Washington University Program on Extremism, 2017), https://extremism.gwu.edu/sites/g/files/zaxdzs5746/files/DigitalDecayFinal_0.pdf, accessed March 18, 2023.
[18] Bennett Clifford and Helen Powell, *Encrypted Extremism: Inside the English-Speaking Islamic State Ecosystem on Telegram*, (Washington DC: George Washington University Program on Extremism, 2019), https://scholarspace.library.gwu.edu/work/9s161692z, accessed March 18, 2023.
[19] Maura Conway *et al*, 'A Snapshot of the Syrian Jihadi Online Ecology: Differential Disruption, Community Strength, and Preferred Other Platforms,' *Studies in Conflict and Terrorism* (2020), https://doi.org/10.1080/1057610X.2020.1866736; Stephane J. Baele, Lewys Brace and Travis G. Coan, 'Uncovering the Far-Right Online Ecosystem: An Analytical Framework and Research Agenda,' *Studies in Conflict & Terrorism* (2020), https://doi.org/10.1080/1057610X.2020.1862895.
[20] 'Telegram FAQ', https://telegram.org/faq, accessed February 9, 2023.
[21] ibid.
[22] Dave Johnson, 'What is Telegram? A quick guide to the fast and secure messaging platform' *Business Insider*, March 24, 2021 https://www.businessinsider.com/what-is-telegram?r=US&IR=T, accessed March 18, 2023.
[23] 'Channels FAQ', https://telegram.org/faq_channels, accessed February 9, 2023.
[24] ibid.
[25] Asaad Almohammad and Charlie Winter, *From Battlefront to Cyberspace: Demystifying the Islamic State's Propaganda Machine*, (West Point, NY: Combating Terrorism Center, 2019), https://ctc.usma.edu/wp-content/uploads/2019/05/Battlefront-to-Cyberspace.pdf, accessed March 18, 2023; Laurence Bindner and Raphael Gluck, 'Assessing Europol's Operation Against ISIS' Propaganda: Approach and Impact', https://icct.nl/publication/assessing-europols-operation-against-isis-propaganda-approach-and-impact/, accessed February 9, 2023.

2. Context

It is then acquired by pro-IS users, following which the dissemination process 'becomes rapidly decentralized'.[26] These users store each piece of propaganda on multiple file-sharing sites, creating large banks of URLs by generating multiple URLs for each item on each site.[27] Often, these file-sharing sites are small or micro companies. A popular example is JustPaste.it. Owned by Mariusz Zurawek, who runs the site out of his home in Poland, Justpaste.it is a free content-sharing service that allows content to be posted within seconds with no registration required. Zurawek receives a large volume of take-down requests from all over the world.[28] This poses challenges in terms of identifying what content is legal and responding to take-down requests in other languages, as well as capacity and resources.

These banks of URLs are then made openly available on public Telegram channels.[29] From here, IS sympathisers can gather the URLs and post them on 'beacon' platforms, such as Twitter.[30] These Twitter ***ghazwah*** (invasions) commonly rely on the use of throwaway accounts, created for the specific purpose of disseminating propaganda and in the expectation that they will be swiftly suspended.[31] The volume of URLs and speed with which they are disseminated are key, often achieved by the use of bots, along with other tactics such as hashtag hijacking and use of the @reply and @mention functions to try and maximise exposure.[32]

In terms of content moderation, Telegram draws a sharp distinction between public and private channels. Its Terms of Service state that, by signing up to Telegram, users agree not to 'Promote violence on *publicly* viewable Telegram channels, bots, etc'.[33] Telegram has in the past taken part in Referral Action Days organised by Europol's EU Internet Referral Unit[34] and, in the first four months of 2022, it claimed to have removed 90,349 terrorist bots and channels.[35] Whilst some have nonetheless doubted Telegram's commitment to moderating publicly available content,[36] its stated approach to public channels stands in marked contrast to its refusal to moderate the contents of private channels, undertaking to 'ensure that no single government or block of like-minded countries can intrude on people's privacy and freedom of expression'.[37] At the same time, Telegram recognises that some users may seek to exploit its public-private dichotomy, stating that 'private channels with publicly available invite links

[26] Daniel Milton, *Pulling Back the Curtain: An Inside Look at the Islamic State's Media Organization*, (West Point, NY: Combating Terrorism Center, 2018), https://ctc.usma.edu/wp-content/uploads/2018/08/Pulling-Back-the-Curtain.pdf, accessed March 18, 2023, 10.

[27] Ahmad Shehabat and Teodor Mitew, 'Black-boxing the black flag: anonymous sharing platforms and ISIS content distribution tactics,' *Perspectives on Terrorism 12, no. 1* (2018): 81-99.

[28] Steven Stalinsky and R. Sosnow, 'The jihadi cycle on content-sharing web services 2009–2016 and the case of Justpaste.it: favored by ISIS, Al-Qaeda, and other jihadis for posting content and sharing it on Twitter – jihadis move to their own platforms (Manbar, Nashir, Alors.Ninja) but then return to Justpaste.it', *MEMRI Inquiry & Analysis Series No* 1255, June 6, 2016, https://www.memri.org/reports/jihadi-cycle-content-sharing-web-services-2009-2016-and-case-justpasteit-favored-isis-al, accessed February 11, 2023.

[29] Stuart Macdonald, Connor Rees and Joost S, *Remove, Impede, Disrupt, Redirect: Understanding & Combating Pro-Islamic State Use of File-Sharing Platforms*, (Washington DC: RESOLVE Network, 2022), https://doi.org/10.37805/ogrr2022.1.

[30] Ali Fisher, Nico Prucha, and Emily Winterbotham, *Mapping the Jihadist Information Ecosystem: Towards the Next Generation of Disruption Capability*, (London: Royal United Services Institute, 2019), https://static.rusi.org/20190716_grntt_paper_06.pdf, accessed March 18, 2023.

[31] Daniel Grinnell et al., *Who disseminates* Rumiyah? *Examining the relative influence of sympathiser and non-sympathiser Twitter users*, https://www.europol.europa.eu/cms/sites/default/files/documents/dgrinnell_smacdonald_dmair_nlorenzodus_who_disseminates_rumiyah_0.pdf, accessed February 11, 2023.

[32] Mohammed Al Darwish, 'From Telegram to Twitter: The Lifecycle of Daesh Propaganda Material', VOX-Pol Blog, September 11, 2019, https://www.voxpol.eu/from-telegram-to-twitter-the-lifecycle-of-daesh-propaganda-material/, accessed February 11, 2023; Macdonald, Rees and S, n 29 above.

[33] 'Terms of Service', https://telegram.org/tos, accessed February 9, 2023 (emphasis added).

[34] 'Europol and Telegram take on terrorist propaganda online,' Europol, https://www.europol.europa.eu/media-press/newsroom/news/europol-and-telegram-take-terrorist-propaganda-online, accessed February 9, 2023.

[35] 'ISIS Watch', https://t.me/s/ISISwatch, accessed February 9, 2023.

[36] Hannah Gais and Megan Squire, 'How an Encrypted Messaging Platform is Changing Extremist Movements', Southern Poverty Law Center, February 16, 2021, https://www.splcenter.org/news/2021/02/16/how-encrypted-messaging-platform-changing-extremist-movements, accessed February 9, 2023.

[37] 'Telegram FAQ', n 20 above.

2. Context

will be treated in the same way as public channels, should it come to content disputes'.[38]

This use of 'aggregator' platforms like Telegram in combination with file-sharing sites and beacon platforms is not limited to IS, nor to jihadist groups more generally. For example, before the Christchurch attacks the attacker uploaded his manifesto to a range of smaller file-sharing sites (including MediaFire, ZippyShare and Solidfiles). Shortly before the first attack, he went onto Facebook, Twitter and 8chan and posted links to the copies of his manifesto available on these file-sharing sites. The post on 8chan also included a link to his Facebook profile, through which he livestreamed the attack. Facebook has reported that the video was viewed fewer than 200 times during the live broadcast. Around this time 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 arrived 12 minutes after the live broadcast ended. The video was subsequently shared on YouTube, as well as the smaller platforms LiveLeak, BitChute and Kiwifarms, and as a downloadable file on Torrentz. Further links to the attack were re-shared on Facebook, Reddit, and 8chan. Whilst most of the smaller platforms reacted responsibly, some did not and did not deactivate links to the video and manifesto.[39]

Facebook has stated that, in the 24 hours after the attacks, it blocked more than 1.2 million videos of the attack at upload.[40] A further 300,000 copies were removed after they were posted. One of the reasons why these additional copies were not detected by Facebook's image and video matching technology was the proliferation of different variants of the video: more than 800 'visually-distinct variants' were in circulation.[41] Some of these were the product of 'a core community of bad actors working together to continually re-upload edited versions of this video in ways designed to defeat our detection'.[42]

Key issues:

- A holistic strategy must address terrorist and extremist exploitation of the variety of online services
- Websites play an important role in online terrorist and extremist ecosystems, yet there is currently a lack of a mitigation strategy
- Propaganda dissemination strategies are underpinned by the use of (often small or micro) file-sharing platforms as repositories for content, many of which lack the capacity or willingness to regulate the content on their platforms

[38] 'Channels FAQ', n 23 above.
[39] Tech Against Terrorism, 'Analysis: New Zealand attack and the terrorist use of the internet', https://www.techagainstterrorism.org/2019/03/26/analysis-new-zealand-attack-and-the-terrorist-use-of-the-internet/, accessed February 9, 2023.
[40] Guy Rosen, 'A Further Update on New Zealand Terrorist Attack', https://about.fb.com/news/2019/03/technical-update-on-new-zealand/, accessed February 9, 2023.
[41] ibid.
[42] ibid.

3. Industry responses

There are two main methods for the identification of TVEC: referrals; and, proactive detection. After outlining each of these, this section then discusses two collaborative initiatives – the Global Internet Forum to Counter Terrorism and Tech Against Terrorism – and the progress to date of each.

a. Referrals

Many platforms offer users the ability to refer content that is believed to violate the terms of service. Users of Twitter, Facebook and TikTok can report tweets, posts and videos. Other platforms have similar mechanisms. For example, Pinterest and Telegram also have 'Report' buttons, and Telegram has an additional email address for takedown requests. Alongside its referral mechanism for individual users, YouTube also has a trusted flagger programme. Now open only to government agencies and NGOs, referrals from trusted flaggers are given priority. Trusted flaggers complete occasional training and are expected to report content with a high accuracy rate. They are also invited to participate in discussion about YouTube content areas.[43]

Another source of referrals is law enforcement. Police forces in several countries have established specialist units, who work to identify TVEC online and refer it to the host platform for removal.[44] In the UK, the Counter Terrorism Internet Referral Unit (CTIRU) was established in 2010. It sits within the Metropolitan Police's Counter Terrorism Command and, during its first eight years, contributed to the removal of 310,000 pieces of content.[45] Following the CTIRU model, the EU's Internet Referral Unit (EU IRU) was established in 2015.[46] Europol describes cooperation with tech companies as a strategic priority, the aim being to exchange best practices and specific measures to improve the referral process and content moderation.[47] One example of cooperation is EU IRU Referral Action Days, which have been organised in collaboration with various companies including SoundCloud,[48] Internet Archive,[49] Telegram,[50] Google,[51] and Facebook.[52]

In the past decade, there has been significant progress in building cooperation between law enforcement and tech companies.[53] But there remain some important challenges. Law enforcement express frustration at the length of time that it can take for requests to be resolved, likening this to a process of 'negotiation'.[54] Meanwhile, the tech sector has raised concerns about the referrals they receive from law enforcement. Sometimes these are only tenuously connected to terrorism, or not connected to it at all. And, while there have been improvements in transparency reporting from the tech sector, there is a feeling that this hasn't been matched by law enforcement or government.[55] These two problems appear to be inter-related: tech companies' follow-up requests for information and justification that follow

[43] 'About the YouTube Trusted Flagger programme', https://support.google.com/youtube/answer/7554338?hl=en-GB, accessed February 11, 2023.
[44] 'Zoey Reeve, 'Repeated and Extensive Exposure to Online Terrorist Content: Counter-Terrorism Internet Referral Unit Perceived Stresses and Strategies', *Studies in Conflict & Terrorism* (2020), https://doi.org/10.1080/1057610X.2020.1792726.
[45] "Together we're tackling online terrorism,' Counter Terrorism Policing, https://www.counterterrorism.police.uk/together-were-tackling-online-terrorism/, accessed February 11, 2023. Members of the public can also report content to CTIRU, including via its iREPORTit app.
[46] "EU Internet Referral Unit - EU IRU: Monitoring terrorism online,' Europol, https://www.europol.europa.eu/about-europol/european-counter-terrorism-centre-ectc/eu-internet-referal-unit-eu-iru, accessed February 11, 2023.
[47] "2021 EU Internet Referral Unit Transparency Report', Europol, https://www.europol.europa.eu/cms/sites/default/files/documents/EU_IRU_Transparency_Report_2021.pdf, accessed February 11, 2023.
[48] "Terrorist and extremist chants used to woo recruits – focus of latest Europol Referral Action Day', Europol, https://www.europol.europa.eu/media-press/newsroom/news/terrorist-and-extremist-chants-used-to-woo-recruits-%E2%80%93-focus-of-latest-europol-referral-action-day, accessed February 11, 2023.
[49] "Jihadist content targeted on Internet Archive platform', Europol, https://www.europol.europa.eu/media-press/newsroom/news/jihadist-content-targeted-internet-archive-platform, accessed February 11, 2023.
[50] "Europol and Telegram take on terrorist propaganda online', n 34 above.
[51] "EU law enforcement and Google take on terrorist propaganda in latest Europol Referral Action Days', Europol, https://www.europol.europa.eu/media-press/newsroom/news/eu-law-enforcement-and-google-take-terrorist-propaganda-in-latest-europol-referral-action-days, accessed February 11, 2023.
[52] "EU law enforcement joins together with Facebook against online terrorist propaganda', Europol, https://www.europol.europa.eu/media-press/newsroom/news/eu-law-enforcement-joins-together-facebook-against-online-terrorist-propaganda, accessed February 11, 2023.
[53] 'Macdonald and Staniforth, n 2 above.
[54] 'Ibid, 14.
[55] 'Ibid.

and slow the response to some referrals seems to be a product of the informality of the process and wider concerns about mission creep. To address this, a 'Takedown-Shutdown Counter Terrorism Policing Protocol' has been proposed, to provide greater transparency, clearly defined referral parameters, and independent oversight for takedown and shutdown requests. [56]

b. Proactive detection

On the biggest social media platforms, referrals account for only a very small proportion of takedowns. On Facebook, the proportion of terrorism-promoting content that is detected proactively, before being reported by users, is roughly 98%.[57] The proactive detection rate on YouTube and Twitter is also above 90%.[58] Unsurprisingly, given the sheer volume of content posted on social media each day, proactive detection relies heavily on AI. Four of the techniques employed by Facebook are: image matching (checking whether a photo or video that is being uploaded to the platform matches a photo or video that has previously been removed for promoting terrorism); language understanding (analysing text that has been removed for promoting terrorism in order to train algorithms to detect similar posts in the future); removing terrorist clusters (using algorithms to work out from groups, posts or profiles that have been identified as supporting terrorism to find other, similar material); and, recidivism (detecting new, fake accounts created by repeat offenders).[59] A triaging process is employed, in which automated systems flag content for humans to review and human judgements are then fed back into the automated systems.[60]

As explained above, volume and speed are key features of propaganda dissemination strategies. This means that behavioural cues are often sufficient to detect TVEC, such as the age of an account, abnormal posting volume and tagging a post with numerous trending hashtags. Cues such as these can be picked up with relative ease by automated systems, meaning such an approach is scalable and often will not require any human intervention. On the other hand, most platforms do not have the resources to build automated content-removal systems. Moreover, when automated systems are used it is important that users have the opportunity to appeal so that a human expert can review potential false positives.[61]

In contrast to behaviour-based decisions, content-based decisions do rely heavily on human involvement. Machines work with data and code; they do not attribute meaning.[62] Contextual nuances such as coded language and irony are better judged by humans. Human expertise is also needed to identify adversarial shifts, where terrorists adapt their strategies in response to and in order to circumvent detection systems.[63] So, even with the development of AI-based tools for detecting TVEC, human decision-making remains essential.

It is not only the smallest companies that lack the necessary capacity for human review.[64] There has been considerable criticism of the size of content moderation teams at the biggest social media

[56] Ibid.
[57] 'Community Standards Enforcement Report – Dangerous Organizations: Terrorism and Organized Hate', n 4 above.
[58] 'YouTube Community Guidelines Enforcement', n 4 above; 'Rules Enforcement', n 4 above.
[59] Monika Bickert and Brian Fishman, 'Hard Questions: How We Counter Terrorism', https://about.fb.com/news/2017/06/how-we-counter-terrorism/, accessed February 11, 2023.
[60] Isabelle van der Vegt et al., *Shedding Light on Terrorist and Extremist Content Removal*, https://gnet-research.org/wp-content/uploads/2019/12/3.pdf, accessed February 11, 2023.
[61] Ibid.
[62] Mireille Hildebrandt, 'Law as computation in the era of artificial legal intelligence: speaking law to the power of statistics,' *University of Toronto Law Journal* 68, supplement 1 (2018): 12–35.
[63] van der Vegt et al., n 60 above.
[64] Hall, n 1 above.

companies, as well as their working conditions – with *the Wall Street Journal* describing it as 'the worst job in technology'.[65] Facebook has a total of 15,000 content moderators, while there is a team of 10,000 to moderate YouTube and other Google products and 1,500 moderators at Twitter.[66] The size of these teams has been described as 'grossly inadequate', particularly given these countries' global coverage and the plethora of national and local languages and cultures.[67] Moreover, the vast majority of the work is outsourced, meaning that most moderators are not employed by the companies themselves.[68] Working conditions are often chaotic, with insufficient time to consider difficult decisions, and 'the peripheral status of moderators undercuts their receiving adequate counseling and medical care for the psychological side effects of repeated exposure to toxic online content'.[69] One study recommended that Facebook double its number of content moderators and bring outsourcing to an end, to allow more time for difficult decisions and greater rotation to protect mental health, while also ensuring a dedicated office in every country in which Facebook does business.[70]

c. Collaborative initiatives

Terrorists' use of a variety of different online services – often in a combined way – means that collaborative initiatives are essential. Perhaps the most prominent example is the Global Internet Forum to Counter Terrorism (GIFCT). Founded by Facebook, Twitter, YouTube and Microsoft in 2017, GIFCT is an NGO with a current total of 22 members. Its activities include the development of cross-platform technical solutions. Its leading initiative is its hash-sharing database. A hash is a numerical representation of a video, image or PDF (akin to a digital fingerprint).[71] When a GIFCT member company removes TVEC, it can create a hash and add it to the shared database. In the event that a user attempts to upload that same item to the platform of another GIFCT member company, the item will automatically be flagged for review. This prevents terrorists jumping from one platform to another, without user data being shared between companies. There are currently 2.1 million hashes in the database, relating to approximately 370,000 unique items of content.[72]

One of the questions addressed in BSR's 2021 Human Rights Impact Assessment of GIFCT was whether GIFCT should actively seek to increase its membership. Stating that two United Nations Guiding Principles on Business and Human Rights (UNGPs) emphasise the importance of prioritising the most severe impacts, BSR concluded that 'GIFCT will be better positioned to prevent terrorists and violent extremists from exploiting digital platforms through more engagement with companies (and organizations) outside the US and Europe, rather than less'.[73] In respect of GIFCT's requirement that all its members publicly commit to respect human rights in accordance with the UNGPs, BSR observed that 'in reality these criteria can be subject to local realities outside of the companies' own control—some companies may, for example, be under local legal expectations to provide direct access to law enforcement agencies or may be partially owned or

[65] Lauren Weber and Deepa Seetharaman, 'The Worst Job in Technology: Staring at Human Depravity to Keep It Off Facebook' *The Wall Street Journal*, December 27, 2017, https://www.wsj.com/articles/the-worst-job-in-technology-staring-at-human-depravity-to-keep-it-off-facebook-1514398398, accessed February 12, 2023.
[66] Paul M. Barrett, *Who Moderates the Social Media Giants? A Call to End Outsourcing*, (New York, NY: NYU Stern Center for Business and Human Rights, 2020), https://issuu.com/nyusterncenterforbusinessandhumanri/docs/nyu_content_moderation_report_final_version?fr=sZWZmZjI1NjI1Ng, accessed February 11, 2023.
[67] Ibid, 2.
[68] Natasha Bernal, 'Facebook's content moderators are fighting back', *Wired*, June 11, 2021, https://www.wired.co.uk/article/facebook-content-moderators-ireland, accessed February 12, 2023; Cristina Criddle, 'Facebook moderator: "Every day was a nightmare"', *BBC News*, May 12, 2021, https://www.bbc.co.uk/news/technology-57088382, accessed February 12, 2023.
[69] Barrett, n 66 above, 1.
[70] Ibid.
[71] https://gifct.org/hsdb, accessed February 12, 2023.
[72] *2022 GIFCT Transparency Report*, https://gifct.org/wp-content/uploads/2022/12/GIFCT-Transparency-Report-2022.pdf, accessed February 12, 2023.
[73] BSR, *Human Rights Assessment: Global Internet Forum to Counter Terrorism*, https://gifct.org/wp-content/uploads/2021/07/BSR_GIFCT_HRIA.pdf, accessed February 12, 2023, 52.

3. Industry responses

controlled by a government associated with human rights harms or complicit in terrorist and violent extremist content activities'.[74] It accordingly proposed a tiered membership structure, with companies initially joining as observers and receiving mentorship from Tech Against Terrorism, and a category of associate membership in which companies would be able to access the hash-sharing database but not add to it. BSR's other recommendations included: extend GIFCT membership beyond just companies operating internet platforms and services to those elsewhere in the tech stack including, in the first instance, those that engage with content issues, such as cloud services companies and content delivery networks; and, extending the technical assistance GIFCT provides to smaller companies to include additional elements relevant to human rights risks, such as the ability to publish transparency reports and to receive and act upon user appeals about content decisions. [75]

Tech Against Terrorism is a private-public partnership backed by the United Nations Counter Terrorism Executive Directorate.[76] Its mentorship programme supports tech companies in meeting GIFCT's membership criteria through knowledge-sharing and capacity-building, including assistance with transparency reporting and understanding how to embed human rights considerations. Tech Against Terrorism also hosts the Terrorist Content Analytics Platform (TCAP), a database of verified terrorist content collected in real-time from messaging platforms and apps.[77] Once content has been added to the TCAP and verified, companies that have registered for the service are sent an automated notification if the content is on their platform. To date, TCAP has identified 38,032 URLs containing terrorist content and sent 21,235 alerts to a total of 73 different companies.[78] Tech Against Terrorism has also recently announced that it is working with Google Jigsaw to build a new prioritisation tool that will ingest the URLs generated by the TCAP and help smaller companies decide how best to manage the large numbers of referrals they receive.[79]

Key issues:

- Tech companies' response to law enforcement takedown and shutdown requests can be delayed by concerns about the content of such requests and the process by which they are made.
- A large volume of TVEC can be detected by automated systems using behavioural cues, but most companies do not have the resources to build such systems.
- Human review remains essential, yet companies of all sizes currently do not have adequate capacity.
- There are promising collaborative initiatives that need to be upscaled, including greater geographic coverage.

[74] Ibid, 53.
[75] Ibid, 57.
[76] https://www.techagainstterrorism.org/, accessed February 12, 2023.
[77] https://www.terrorismanalytics.org/, accessed February 12, 2023.
[78] 'Tech Against Terrorism to Build Content Moderation Tool with Google Jigsaw', January 9, 2023, https://www.techagainstterrorism.org/2023/01/09/tech-against-terrorism-to-build-content-moderation-tool-with-google-jigsaw/, accessed February 12, 2023.
[79] Ibid.

4. Discussion

The following discussion focuses on four sets of issues: transparency; definitional clarity; the impact on those targeted; and, the use of online data for predictive purposes. The premise underlying the discussion is that, while tech companies do not have the obligations of governments, their function and impact means that they should respect human rights standards. Indeed, one of the criteria for membership of the Global Internet Forum to Counter Terrorism (GIFCT) is a public commitment to human rights, in accordance with the United Nations Guiding Principles on Business and Human Rights.

a. Transparency

The importance of transparency has been emphasised in numerous different settings, with reasons including preventing corruption, uncovering mistakes, building trust, improving public debate, enhancing democracy and promoting accountability.[80] In the current context, the focus has been largely on two transparency mechanisms: the publication of content moderation policies; and, publicly available reports containing statistical data and breakdowns.[81] Each of these is a criterion for membership of both GIFCT and Tech Against Terrorism, who emphasise the importance of transparency in promoting multi-stakeholder collaboration, as well as sharing learning, correcting misunderstandings and enhancing accountability.[82]

Relevant EU legislation imposes transparency requirements. The Terrorist Content Online Regulation requires hosting service providers (of all sizes) to publish an annual transparency report and the company's policy to prevent the dissemination of terrorist content, including the details of any automated tools.[83] Alongside these obligations, the EU's Internal Security Fund work programme for 2021/22 includes funding for activities to support small tech companies in implementing the Regulation. Transparency reporting requirements will be strengthened by the EU's Digital Services Act. For all but small and micro platforms,[84] the Act requires annual, publicly available, easily comprehensible reports containing: information on content moderation policies and practices; details of any use made of automated means for the purpose of content moderation; and, data on orders received from authorities in Member States, referral notices and complaints received and responses to these, among other things.

The UK's Online Safety Act will also create formal transparency requirements.[85] Service providers will be required to produce annual transparency reports for each of their services, with OFCOM determining the information to be included in these reports in a notice given to the provider. Schedule 8 of the Act lists the matters about which information may be required. It also stipulates that, when determining which information should be required in a notice, OFCOM must take into account the number of users of the service, the capacity of the provider, and the proportion of users who are children, among other things.

Concerns about current transparency reporting practices include: selective use of metrics; lack of contextual information to enable a full understanding of the data provided; the use of proportional metrics that fail to give an accurate indication of the scale of harm; and, the difficulty in making cross-platform comparisons when companies use different metrics.[86] A further concern is the lack of access for independent researchers. Such access, which is necessary for independent evaluation and validation of internal

[80] Elizabeth Fisher, 'Transparency and Administrative Law: A Critical Evaluation', *Current Legal Problems*, 63, no. 1, (2010): 272-314.
[81] Courtney Radsch, *Transparency Reporting: Good Practices and Lessons from Global Assessment Frameworks*, https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-ResearchAgendaScopingPaper-1.1.pdf, accessed February 16, 2023.
[82] BSR, n 73 above.
[83] Regulation 2021/784, Article 7.
[84] Regulation 2022/2065, Articles 15 and 42. According to Directive 2003/361/EC, a small enterprise is defined as one which employs fewer than 50 persons and whose annual turnover and/or annual balance sheet total does not exceed €10 million. For micro enterprises the figures are ten staff and €2 million.
[85] Sections 77-78.
[86] Joint Committee on the Draft Online Safety Bill, *Draft Online Safety Bill*, Report of Session 2021-22, https://committees.parliament.uk/publications/8206/documents/84092/default/, accessed February 16, 2023.

company studies,[87] has been opposed for reasons including user privacy.[88] The present lack of access 'hinders much-needed scientific progress towards understanding the prevalence, impact, causes, and dynamics of online activity that creates a risk of harm'.[89] The Digital Services Act will impose a requirement on providers of very large online platforms and search engines to provide access to vetted researchers for 'the sole purpose of conducting research that contributes to the detection, identification and understanding of systemic risks in the Union'.[90] The Online Safety Act requires OFCOM to publish a report on researchers' access to data within 18 months of the Act's enactment.[91] The Joint Committee on the Act also recommended that it requires service providers to conduct risk assessments of opening up data on online safety to independent researchers, including the impact on privacy.[92]

b. Definitional clarity

Definitional clarity is important for several reasons. It provides users with fair warning of what content is not permissible, enabling them to make informed decisions about their use of the platform.[93] It limits the discretion of content moderators, ensuring greater consistency in decision-making while guarding against potential misuse of power and censorship creep.[94] It also helps ensure that users have an effective opportunity to appeal moderation decisions, should their content be taken down.[95]

The difficulties of defining terrorism are well-known. Concocting a legal definition of terrorism has been described as a trilemma: adopt an under-inclusive definition that excludes all attacks on the state and its officials; adopt an over-inclusive definition that encompasses legitimate freedom fighters; or, adopt a definition that discriminates between legitimate and illegitimate attacks on the state and, in so doing, requires legal actors to make political judgments that they have inadequate expertise to make.[96]

In terms of the moderation of online TVEC, there are three factors that further exacerbate the definitional complexities. First, there is the question whether it is appropriate for tech companies to be determining the parameters of permissible speech. There are concerns here about the companies' moral legitimacy, accountability deficits and augmenting the power of the powerful.[97] The UN Special Rapporteur on the right to freedom of opinion and expression has stated that governments should 'avoid delegating responsibility to companies as adjudicators of content, which empowers corporate judgment over human rights values to the detriment of users'.[98]

Second, in practice the definition of terrorism will either be applied by human moderators, working in the conditions described above, or by automated systems. There have been a number of examples of automated systems erroneously removing content for violating policies on TVEC, including materials providing evidence of human rights violations.[99]

[87] Mark MacCarthy, 'Transparency is essential for effective social media regulation', https://www.brookings.edu/blog/techtank/2022/11/01/transparency-is-essential-for-effective-social-media-regulation/, accessed February 16, 2023.
[88] Joint Committee on the Draft Online Safety Bill, n 86 above.
[89] Ibid, 120.
[90] Article 40. 'Very large' is defined as more than 45 million average monthly users of the service in the EU (Article 33).
[91] Section 162.
[92] Joint Committee on the Draft Online Safety Bill, n 86 above, 123. The Independent Reviewer of Terrorism Legislation has also recommended that counterterrorism police create and publish a list of content whose possession or dissemination has led to convictions in the UK under section 58 of the Terrorism Act 2000 and section 2 of the Terrorism Act 2006. One benefit of such a list would be to assist tech companies with content moderation decisions (Hall, n 1 above).
[93] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/HRC/38/35, https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/096/72/PDF/G1809672.pdf?OpenElement, accessed February 16, 2023.
[94] Danielle Citron 'Extremist Speech, Compelled Conformity, and Censorship Creep', *Notre Dame Law Review*, 93, no. 3 (2018): 1035-1072; Jeffrey Howard, 'Should we ban dangerous speech? *British Academy Review*, 32 (2018): 19-21.
[95] Stuart Macdonald, Sara Giro Correia and Amy-Louise Watkin, 'Regulating terrorist content on social media: automation and the rule of law', *International Journal of Law in Context*, 15, no. 2 (2019): 183-197.
[96] Jacqueline S. Hodgson and Victor Tadros, 'The impossibility of defining terrorism', *New Criminal Law Review*, 16, no. 3, (2013): 494–526.
[97] Alastair Reed and Adam Henschke, 'Who Should Regulate Extremist Content Online?' in Adam Henschke, Alastair Reed, Scott Robbins and Seumas Miller (Eds.), *Counter-Terrorism, Ethics and Technology: Emerging Challenges at the Frontiers of Counter-Terrorism* (Cham: Springer, 2021), 175-198; Evelyn Douek, 'The Rise of Content Cartels', Knight First Amendment Institute at Columbia University, https://knightcolumbia.org/content/the-rise-of-content-cartels, accessed February 16, 2023.
[98] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, n 93 above.
[99] Macdonald, Correia and Watkin, n 95 above.

4. Discussion

Third, collaborative initiatives like the GIFCT hash-sharing database will be most effective if there is consensus as to the scope of prohibitions on TVEC. Yet across GIFCT member companies there is no common approach to defining terrorist content.[100] While the human rights impact assessment of GIFCT's strategy, governance and operations stopped short of recommending the adoption of a shared definition, it did recommend the development of a 'common understanding'.[101]

Defining violent extremism is an equally complex task. The UN Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism has suggested that the term extremism is 'conceptually weaker than the term terrorism, which has an identifiable core'.[102] This can leave companies open to pressure from government authorities to remove content on questionable grounds.[103] Research is needed to better understand how the term is applied in practice, in particular, whether content is removed as being violent extremist that would not otherwise violate prohibitions on terrorist and hateful or violent content.[104]

c. The impact on those targeted

There is evidence that suggests that far-right TVEC is less likely to be removed than jihadist content. Tech Against Terrorism's analysis of responses to its TCAP alerts (which all related to official materials from designated terrorist entities) found a significantly higher takedown rate for jihadist content (94%) than far-right content (50%).[105] Various reasons were suggested for this. The branding used by jihadist groups may be more readily recognised by non-experts in tech companies than the symbols used by far-right groups. The platforms on which far-right content is hosted often have a higher threshold for removal, which some platforms seek to justify by reference to the First Amendment to the US Constitution. There could also be jurisdictional reasons, such as where a US company is asked to remove content produced by an organisation that is proscribed in the UK but not the US. In such a situation, the company might only remove the content for users in a particular jurisdiction, leaving it still accessible by users in that jurisdiction using a VPN.[106]

Stronger enforcement action against jihadist groups than far-right ones has the potential to be perceived as discriminatory. This is particularly important in the current context, given that claims of anti-Muslim prejudice are utilised by jihadist radicalisers who deploy an us versus them discourse to Other the West.[107] Indeed, one study of IS activity on Twitter found that suspension played an important role in community-building, with the majority of the accounts studied referring to Twitter's use of suspension as a specific tool to persecute Muslims.[108]

Algorithmic decision-making can also impact different groups of individuals differently, for example, as a result of non-representative data collection.[109] For this reason, it is important to examine the actual outcomes of algorithmic decisions.[110] The Digital Services

[100] Katy Vaughan, The Interoperability of Terrorism Definitions (Washington, DC: GIFCT, 2022), https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-LF-TVEC-1.1.pdf, accessed February 16, 2023.
[101] BSR, n 73 above, 35.
[102] Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, A/HRC/40/52, https://documents-dds-ny.un.org/doc/UNDOC/GEN/G19/057/59/PDF/G1905759.pdf?OpenElement, accessed February 16, 2023, 11.
[103] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, n 93 above.
[104] Vaughan, n 100 above.
[105] *Transparency Report: Terrorist Content Analytics Platform, Year One: 1 December 2020 – 30 November 2021*, n 6 above.
[106] 'Comparative Analysis of the TCAP Transparency Report Statistics on Content Collection and Removal Rates', https://terrorismanalytics.org/project-news/comparative-analysis-of-the-tcap-transparency-report, accessed February 17, 2023.
[107] Nuria Lorenzo-Dus and Stuart Macdonald, 'Othering the West in the online jihadist propaganda magazines *Inspire and Dabiq*', *Journal of Language, Aggression and Conflict*, 6, no. 1, (2018): 79–106.
[108] Elizabeth Pearson, 'Online as the new frontline: affect, gender, and ISIS-take-down on social media', *Studies in Conflict & Terrorism*, 41, no. 11 (2018): 850-874.
[109] David Lehr and Paul Ohm, 'Playing with the data: what legal scholars should learn about machine learning', *University of California, Davis, Law Review*, 51, no. 2, (2017): 653–717.
[110] Anupam Chandler, 'The racist algorithm', *Michigan Law Review*, 115, no. 6 (2017): 1023–1045.

Act obliges very large online platforms and search engines to conduct annual, independent audits, including access to all relevant data and premises, to assess their compliance with the obligations imposed by the Act.[111] Under the Online Safety Act, OFCOM will have the power to undertake audits and to require skilled person reports.[112] Key to the effectiveness of algorithmic auditing are the criteria used to assess systems and the procedures used to assess against these criteria.[113] A recent Government discussion paper concluded that, other than in highly regulated sectors, the algorithm audit landscape lacks specific rules and standards. Auditors are also often limited by a lack of access to systems and reluctance on the part of organisations to cooperate.[114]

d. The use of online data for predictive purposes

While terrorists use online platforms in support of their activities, it is also the case that security agencies and law enforcement use the internet to interdict attacks. It has even been suggested that security actors gain at least as much utility from the internet as terrorists do.[115] The internet offers governments enhanced access to information and power to coordinate, as well as the opportunity to gain more information on the terrorists themselves – especially as many terrorists overestimate the level of anonymity they enjoy online.[116] There is some empirical support for this perspective; recent studies have found those that engaged in an online network were far less likely to succeed in their plot than those that did not.[117]

This raises the question whether AI can be trained to use online data to detect and predict terrorist activity. While obviously attractive, such efforts face a number of challenges. The first concerns the datasets used in existing studies on the potential of AI to be used in this way. Datasets collected for these studies are (for understandable reasons) rarely made openly available, which means that they cannot be verified. The datasets could contain false positives (content or user accounts that have been erroneously categorised as terrorist), which would impair the performance of algorithms trained on them.[118] Moreover, the datasets may not be representative of the larger population of interest. For example, datasets that are collected based on selected terms and expressions may only cover the terminology of particular subgroups.[119] There are also methodological problems with existing studies. Many lack any comparison with a control group. When a control group is used, these are often composed of randomly collected posts and user accounts, i.e., ordinary users talking about issues not related to extremism or terrorism. Yet the key challenge is to differentiate extremist accounts from those that – despite using the same terminology, reporting the same events, or talking about the same topics – are not extremist.[120]

[111] Article 37.
[112] HM Government, *Government Response to the Report of the Joint Committee on the Draft Online Safety Bill*, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1061446/E02721600_Gov_Resp_to_Online_Safety_Bill_Accessible_v1.0.pdf, accessed February 17, 2023, 46.
[113] Algorithm Watch, 'Our response to the European Commission's planned Digital Services Act', https://algorithmwatch.org/en/submission-digital-services-act-dsa/#audit, accessed February 17, 2023.
[114] Digital Regulation Cooperation Forum, *Auditing algorithms: the existing landscape, role of regulators and future outlook* (2022), https://www.gov.uk/find-digital-market-research/auditing-algorithms-the-existing-landscape-role-of-regulators-and-future-outlook-2022-drcf, accessed February 17, 2023.
[115] David C Benson, 'Why the Internet is Not Increasing Terrorism', *Security Studies*, 23, no.2 (2014): 293-328.
[116] Ibid.
[117] Joe Whittaker, 'The online behaviors of Islamic state terrorists in the United States', *Criminology & Public Policy*, 20, no. 1 (2021): 177-203: see also Keynon, Binder and Baker-Beall, n 3 above.
[118] Miriam Fernandez and Harith Alani, 'Artificial Intelligence and Online Extremism: Challenges and Opportunities' in John McDaniel and Ken Pease (eds.) *Predictive Policing and Artificial Intelligence* (Abingdon: Routledge, 2021), 132–162.
[119] ibid.
[120] ibid.

4. Discussion

Second, it is difficult to develop generic online radicalisation detection methods when the data comes in multiple languages, from multiple platforms, in multiple formats.[121] More generally, as some margin of error is inevitable, a choice must be made whether to prioritise the reduction of false positives or false negatives. Optimising for false positives would mean more tolerance of relevant users escaping undetected, while optimising for false negatives would mean accepting incorrectly identifying some users as terrorist suspects.[122]

Third, as noted above, algorithms can struggle with more nuanced communication, such as irony and sarcasm.[123] In its study of abuse on Twitter against Premier League footballers, the Alan Turing Institute developed a machine learning tool that automatically assessed whether tweets were abusive. While the tool performed well, this was because it was highly adapted for the specific task at hand, meaning it 'may be brittle to small changes in the setting or task' and so could perform poorly if applied in a different domain.[124] This is significant, since terrorist groups perform so-called adversarial shifts, adapting their behaviour to avoid detection. It is therefore necessary to keep retraining AI tools so that they keep up with this constant evolution.[125]

The upshot is that human expertise remains essential and must be integrated into the decision-making process for AI solutions to be effective.

Key issues:

- It is necessary to improve and harmonise transparency reporting practices and to provide independent researchers with access to data.
- Greater consensus around the meaning of terrorism would enhance collaborative initiatives. Key stakeholders, including governments, should be involved in this process.
- There is a need to understand the practical use and value of the term violent extremism.
- The feeling that enforcement action targets a specific group or community can be exploited by radicalisers.
- Efforts to develop AI that can use online data to detect and predict terrorist activity face several challenges. Human expertise will remain essential and must be kept in the loop in the development of new technology.

[121] ibid.
[122] UNICRI and UNCCT, *Countering Terrorism Online with Artificial Intelligence: An Overview for Law Enforcement and Counter-Terrorism Agencies in South Asia and South-East Asia* (New York, NY: UNOCT, 2021), https://www.un.org/counterterrorism/sites/www.un.org.counterterrorism/files/countering-terrorism-online-with-ai-uncct-unicri-report-web.pdf, accessed February 17, 2023.
[123] ibid.
[124] Bertie Vidgen *et al.*, *Tracking abuse on Twitter against football players in the 2021-22 Premier League Season* (The Alan Turing Institute, 2022), https://www.turing.ac.uk/sites/default/files/2022-08/tracking_abuse_on_twitter_against_football_players_web.pdf, accessed February 17, 2023, 29.
[125] Fernandez and Alani, n 118 above.

5. Conclusion

The use of AI has the potential to improve tech companies' capacity to identify and remove terrorist content. It is possible to use automated systems to detect a large volume of TVEC using behavioural cues. However, most companies do not have the resources to build such systems and existing collaborative initiatives need to be upscaled. Human review also remains essential, both for content-based decisions and for the development of AI that can use online data to detect or predict terrorist activity.

This paper has also identified a number of other, wider issues. The prevalence of terrorist operated websites and concerns about law enforcement takedown and shutdown requests need to be addressed. Respect for human rights also requires an improvement in transparency reporting, including access to data for independent researchers, as well as greater consensus around the meaning of key terms and auditing of the outcomes of algorithmic decision-making.

6. Recommendations

- Governments should develop a **global mitigation strategy to combat terrorist operated websites**, in collaboration with the tech sector. This is key to a holistic approach to tackling online TVEC.

- **A (publicly available) protocol for counterterrorism law enforcement takedown and shutdown requests** should be implemented. This would include clearly defined referral parameters and the introduction of independent oversight for takedown and shutdown requests.

- **Automated systems that use behavioural cues to identify online TVEC** should be developed and made available to those companies that lack the capacity to develop such tools themselves.

- **GIFCT membership needs to be expanded**, including the recruitment of members from elsewhere in the tech stack and from non-US locations. A tiered membership structure could be used to manage the human rights risks of expansion.[126]

- Transparency reporting requirements should ensure that the metrics used enable cross-platform comparisons to be made. A concerted effort is also needed to **provide independent researchers with access to data**. To facilitate this, service providers should conduct and make available risk assessments of providing such access, including measures for mitigating the impact on user privacy.

- There are definitional issues that need to be addressed. Research is needed to better understand the practical operation of the term violent extremism. Meanwhile, collaborative initiatives such as the GIFCT hash-sharing database would be enhanced by the **development of a common understanding of terrorist content**. This common understanding should be drawn as narrowly as possible and be sufficiently granular to be actionable and practical for companies to use.[127]

[126] BSR, n 73 above, 55.
[127] BSR, n 73 above, 35.

7. Further reading

BSR, *Human Rights Assessment: Global Internet Forum to Counter Terrorism*, https://gifct.org/wp-content/uploads/2021/07/BSR_GIFCT_HRIA.pdf

Evelyn Douek, 'Governing Online Speech: From "Posts-As-Trumps" to Proportionality and Probability', *Columbia Law Review*, 121, no. 3 (2021): 759-833.

Isabelle van der Vegt, Paul Gill, Stuart Macdonald and Bennett Kleinberg, *Shedding Light on Terrorist and Extremist Content Removal*, https://gnet-research.org/wp-content/uploads/2019/12/3.pdf

Jonathan Hall, *The Terrorism Acts in 2021: Report of the Independent Reviewer of Terrorism Legislation on the Operation of the Terrorism Acts 2000 and 2006, and the Terrorism Prevention and Investigation Measures Act 2011* (His Majesty's Stationery Office, 2023), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1140911/E02876111_Terrorism_Acts_in_2021_Accessible.pdf.

Katy Vaughan, *The Interoperability of Terrorism Definitions* (Washington, DC: GIFCT, 2022), https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-LF-TVEC-1.1.pdf

Stuart Macdonald, Sara Giro Correia and Amy-Louise Watkin, 'Regulating terrorist content on social media: automation and the rule of law', *International Journal of Law in Context*, 15, no. 2 (2019): 183-197.

Stuart Macdonald and Andrew Staniforth, *Tackling Online Terrorist Content Together: Counterterrorism Law Enforcement and Tech Company Cooperation*, (London: Global Network on Extremism and Technology, 2023), https://gnet-research.org/wp-content/uploads/2023/01/31-Tackling-Online-Terrorist-Content-Together_web.pdf

Tech Against Terrorism, *Responding to Terrorist Operated Websites*, (London: Tech Against Terrorism, 2022), https://www.techagainstterrorism.org/wp-content/uploads/2022/07/TAT-TOW-Mitigation-Strategy-July-2022.pdf

NOTIONES

Coordinator	Scientific Technical Coordinator	Project Security Officer
tecnalia MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE	Zanasi & Partners Security Research and Advisory	

Academic | Think-Tanks | Research

LAUREA · DRI Defence Research Institute · Fondazione ICSA · Bar-Ilan University · APRE Agenzia per la Promozione della Ricerca Europea

Technology Providers

VTT · expert.ai · SAHER EUROPE · MARKETSCAPE · Tecoms security through research · SYNYO

Practitioners

POLICJA MAZOWSZE · Financial Intelligence Unit · ISEM INSTITUTE International Security and Emergency Management Institute · Keeping People Safe · Hochschule für den öffentlichen Dienst in Bayern Fachbereich Polizei · BEYOND THE HORIZON

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 101021853.