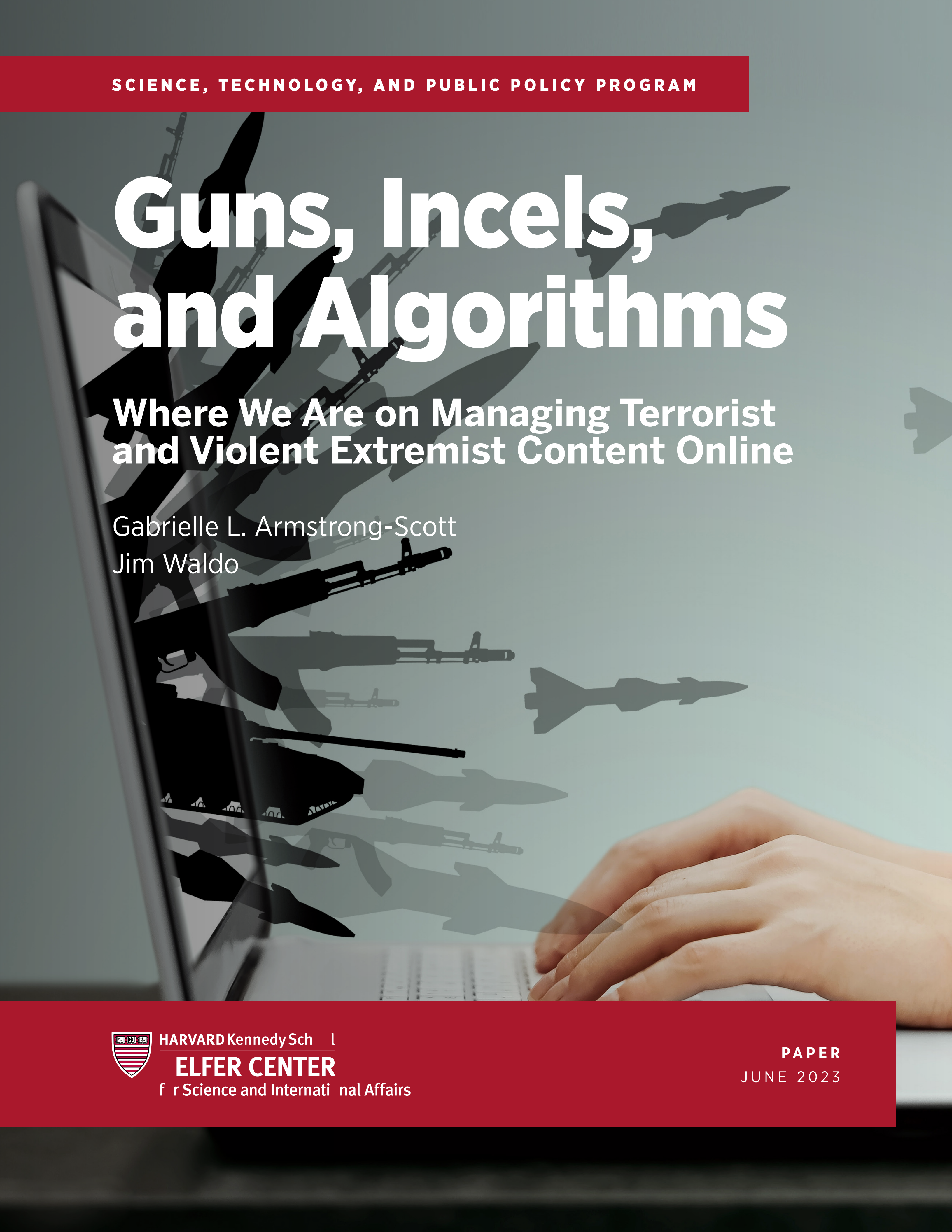
Guns, Incels, and Algorithms: Where We Are on Managing Terrorist and Violent Extremist Content Online



SCIENCE, TECHNOLOGY, AND PUBLIC POLICY PROGRAM

Guns, Incels, and Algorithms

Where We Are on Managing Terrorist and Violent Extremist Content Online

Gabrielle L. Armstrong-Scott
Jim Waldo

HARVARD Kennedy Sch l
ELFER CENTER
f r Science and Internati nal Affairs

PAPER
JUNE 2023

Science, Technology, and Public Policy Program
Belfer Center for Science and International Affairs
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138

www.belfercenter.org/stpp

Statements and views expressed in this report are solely those of the author(s) and do not imply endorsement by Harvard University, Harvard Kennedy School, or the Belfer Center for Science and International Affairs.

SCIENCE, TECHNOLOGY, AND PUBLIC POLICY PROGRAM

Guns, Incels, and Algorithms

Where We Are on Managing Terrorist and Violent Extremist Content Online

Gabrielle L. Armstrong-Scott
Jim Waldo

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs

PAPER
JUNE 2023

About the Program

The Science, Technology, and Public Policy (STPP) Program draws on insights from scholarly and applied work in science and technology, technology assessment, political science, economics, management, and law to research and practice on the intersection of science and technology with public affairs. The goal is to help develop and promote public policies that advance the application of science and technology to improvement of the human condition.

For more, visit belfercenter.org/STPP

About the Authors

Gabrielle (Gabe) Armstrong-Scott is a Graham T. Allison, Jr. Fellow at Harvard Kennedy School's Belfer Center for Science and International Affairs. She previously worked in national security for the New Zealand Department of the Prime Minister and Cabinet. She has also worked at the New Zealand Mission to the UN and as a technology and geopolitical risk consultant in New York and DC. At Harvard, Armstrong-Scott was teaching assistant to former U.S. Secretary of Defense Ash Carter and was a Knox Fellowship recipient, for "future promise of leadership, strength of character, keen mind, balanced judgement and a devotion to the democratic ideal." Armstrong-Scott has an A.B. from Princeton University's School of Public and International Affairs and a M.P.E. (Economics) from Victoria University, where she won the Excellence Prize for top master's student.

James (Jim) Waldo is the Gordon McKay Professor of the Practice of Computer Science in the School of Engineering and Applied Sciences at Harvard, where he teaches courses in distributed systems and privacy; the Chief Technology Officer for the School of Engineering and Applied Sciences; and a Professor of Policy teaching on topics of technology and policy at Harvard Kennedy School. Jim designed clouds at VMware; was a Distinguished Engineer with Sun Microsystems Laboratories, where he investigated next-generation large-scale distributed systems; and got his start in distributed systems at Apollo Computer. While at Sun, he was the technical lead of Project Darkstar, a multi-threaded, distributed infrastructure for massive multiplayer online games and virtual worlds; the lead architect for Jini, a distributed programming system based on Java; and an early member of the Java software organization. Jim is the author of *Java: the Good Parts* (O'Reilly) and co-authored *The Jini Specifications* (Addison-Wesley). He edited *The Evolution of C++: Language Design in the Marketplace of Ideas* (MIT Press). He co-chaired a National Academies study on privacy and co-edited the report "Engaging Privacy and Information Technology in a Digital Age." He is the author of numerous journal and conference proceedings articles, and holds over fifty patents.

Table of Contents

Executive Summary

Ten years ago, U.S. national security agencies grew concerned about a relatively new and powerful weapon used by terrorists: the World Wide Web. What had begun as an effort to connect end users from across the world to share information and to serve as a force of human liberation, instead began to be used as a tool for destruction of life. Terrorists were exploiting technology companies' lax content moderation policies to recruit new members, spread violent extremist ideology, and plan terrorist attacks. In 2012, Twitter's General Manager declared the firm "the free speech wing of the Free Speech Party," and large U.S. technology companies were broadly reticent to make changes to their content moderation policies in the early days of their development.[1]

By 2015, a gargantuan effort to eliminate ISIS commenced – mostly driven by the U.S. government – culminating in U.S. Cyber Command's Operation GLOWING SYMPHONY, led by General Paul Nakasone, which reportedly foiled the majority of ISIS' online presence and networks in 2016. Technology companies became much stricter about terrorist content online, but the problem of identifying and removing such content persisted.[2]

Today, the online terrorism landscape looks much different to a decade ago. White supremacist and "incel" (involuntary celibate) violent extremist content litters the Web. Terrorist attacks are frequently committed by hate-fuelled lone-wolf "internet warriors" who have been inspired by non-Islamic terrorist and violent extremist content and radicalizing material online. Yet, technology companies and governments have not managed to keep pace with the dynamic threat.[3]

This is not to say that they haven't tried. In 2019, a terrorist attack committed (and live-streamed, going viral) by an "online warrior" white supremacist at two mosques in Christchurch, New Zealand, galvanized technology companies and governments to do more to combat terrorist

[1] "GIFCT Working Groups Output 2022."

[2] Temple-Raston, "How The U.S. Hacked ISIS."

[3] Cai and Landon, "Attacks by White Extremists Are Growing. So Are Their Connections."

content beyond just Islamic terrorism, culminating in an ambitious multilateral initiative, *The Christchurch Call to Eliminate Terrorist and Violent Extremist Content Online*, an unprecedented diplomatic achievement and step forward in managing the problem.[4]

Technology companies and governments have spent the past decade trying to better address the evolving threat of terrorist and violent extremist content online (TVEC). However, there are few studies examining just how effective these efforts have been, where we are today in managing the problem, and wherein lie gaps for improvement.

This paper argues that companies' efforts to deal with TVEC have been hampered at the outset by a tendency to define TVEC extremely narrowly. Still, only a tiny proportion of content that could reasonably be categorized as TVEC is included in most definitions. An outsized focus on pre-identified Islamic extremists and terrorist groups means that other types of violent extremists and terrorists (e.g., white supremacists, incels), and those unaffiliated with a group (e.g., lone-wolf actors) are overlooked. This paper also explores the idea of ethical obligations and norms as an alternative to a legally required definition.

On the technical side, this paper finds that even if there was consensus on the legal and ethical questions surrounding TVEC, the technical tools currently available are no panacea. Trade-offs across efficiency, scalability, accuracy, and resilience are persistent. Current technical tools tend to disadvantage minority groups and non-English languages. They are also less robustly implemented across small and non-U.S./European firms, generally either because they are left out of inter-firm initiatives or because they lack resources and capability. This paper does not claim to cover every issue relevant to TVEC; however, it highlights several important gaps that could be addressed by policymakers and tech companies and identifies avenues for future research.

[4] Call, "Christchurch Call Text."

It concludes the following:

1. A uniform and broader definition of TVEC should be formulated by policymakers and implemented across technology companies, to encompass specified *actions* or *activities* and *unaffiliated actors*, beyond just designated Islamic terrorist entities.
2. Not all technical tools are created equal. Multilateral and cross-company initiatives to combat TVEC should be inclusive of smaller firms and non-U.S. and non-European firms.
3. Development of TVEC identification and management tools that are well-trained across different languages and cultural contexts is needed to ensure equitable standards in managing TVEC.
4. A standard of success needs to be established for machine learning (ML) tools to guide progress towards an ideal 'North Star'. As it stands, ML tools are not yet good enough to "algorithm our way out of the problem,"[5] and a combination of tools is required, as no tool yet deals with the full extent of TVEC in all its forms.
5. Legal regimes of corporate social responsibility that emphasize saving lives in the real world by managing TVEC online would liberate technology companies from their obligations to shareholders to engage in practices that maximize engagement and profit at all costs.
6. Policymakers should be wary of unintended consequences of well-intentioned policies, such as relocation to smaller and less-regulated platforms after being de-platformed.
7. Public-private cooperation is critical in managing the threat of TVEC from a national security perspective.

[5] "Challenges in Combating Terrorism and Extremism Online."

1. Introduction

Four years ago, a terrorist murdered 51 people at two mosques in Christchurch, New Zealand. The attack was livestreamed online, and millions of copies of the video of the attack and the terrorist's manifesto were uploaded to mainstream technology platforms within hours.[6] The terrorist himself had been inspired by terrorist and violent extremist content (TVEC), and his video and manifesto became an inspiration for would-be attackers, just as other white supremacist terrorist content online had inspired him.[7] Technology platforms struggled to identify and remove the content, prompting a reckoning among governments and private industry that more needed to be done to address this problem.[8]

A raft of new efforts to counter TVEC emerged from this moment. While some initiatives had already been established following a spate of terrorist incidents in the 2010s and the rise of ISIS' online presence, the level of multilateral cooperation observed in the year following the Christchurch attacks was unprecedented in terms of large technology platforms cooperating for content moderation (the only exception being efforts to eliminate child sexual abuse material). The technology industry appears to have experienced a normative shift since Christchurch towards a greater overall willingness to take measures to counter violent extremist content, albeit not observed completely across the board.

This decade, TVEC's role in fueling real-life terrorism and violent extremism has become much more salient. Until 2015, efforts to counter TVEC were limited, and there was a laissez-faire culture of content moderation among tech platforms.

Even ISIS material – its existence against most social media companies' policies even at the time – faced little tangible mitigation efforts by companies. A 2015 Brookings Institute study found that "…social media companies have for almost a decade facilitated the rapid growth of virtual communities of terrorists and their sympathizers…at the very least, software that recognizes terrorist logos and symbols could be used by social media companies to flag accounts for preliminary

[6] "Facebook Says It Has Removed 1.5 Million Copies of the New Zealand Terror Attack Video."

[7] Cai and Landon, "Attacks by White Extremists Are Growing. So Are Their Connections."

[8] "The Report"; "5 Months on, Christchurch Attacker Influences Others."

review, but this has not yet happened."[9] Two more studies from Recorded Future and Brookings discovered the existence of around 60,000 active pro-ISIS Twitter accounts, despite being against Twitter policy.[10] There arose a recognition from governments and civil society that social media companies needed to do more to counter ISIS content.

Contrary to private industry, the U.S. government played an activist role in countering TVEC in the 2010s: U.S. national security agencies commenced initiatives to spread counter-messaging and operations to identify, surveil, and foil terrorists using their online networks.[11] Among lawmakers in the United States, there was hesitation to exert more control over social media companies for constitutional reasons, though national security agencies were able to pursue then-classified cyber operations to foil ISIS networks. U.S. Cyber Command's Joint Task Force ARES, led by General Paul Nakasone, conducted an offensive cyber operation called Operation GLOWING SYMPHONY through 2015 and 2016 which was reportedly highly successful at eliminating much of ISIS's online networks and social media presence.[12]

U.S. national security agencies and law enforcement continue to take cyber-related measures to foil terrorist groups and potential terrorist threats, and there is a long history and culture in the United States of government-private industry cooperation to limit national security threats in the cyber domain that will almost certainly continue.[13]

However, as terrorist and violent extremist content online has become increasingly diffuse, and many terrorists and violent extremists today are not affiliated with any one group, it becomes harder to predict where would-be terrorists lie compared with the interconnected nature of ISIS networks.[14] It is worth examining whether national security agencies – with demonstrated top-level expertise – could play a larger role in helping social media companies guard against the radicalizing force

[9] Alberto M Fernandez, "Here to Stay and Growing: Combating ISIS Propaganda Networks," Brookings Institute, 2015, pg. 29.

[10] Fernandez; Berger, "The ISIS Twitter Census."

[11] Cottee, "Why It's So Hard to Stop ISIS Propaganda"; Miller and Higham, "In a Propaganda War against ISIS, the U.S. Tried to Play by the Enemy's Rules"; Schmitt, "U.S. Intensifies Effort to Blunt ISIS' Message."

[12] Temple-Raston, "How The U.S. Hacked ISIS."

[13] "FBI Partnering with the Private Sector to Counter the Cyber Threat"; "Innovative Public Private Partnerships"; Stiglitz and Wallsten, "Public-Private Technology Partnerships"; Carr, "Public-Private Partnerships in National Cyber-Security Strategies."

[14] Cai and Landon, "Attacks by White Extremists Are Growing. So Are Their Connections."

of TVEC in forms beyond Islamic extremism today.

From 2016, a spate of terrorist attacks fuelled by online radicalization incentivized European governments to introduce new legal regimes with stringent terms on TVEC regulation. Without a First Amendment – which has heavily constrained legal changes to manage TVEC in the United States – it was simpler for the EU to take this step. The attacks lent ever more credence to the idea that TVEC and its real-life effects were important negative societal implications of the spread of social media, instant messaging tools and the general explosion of Web usage into the billions.[15]

As TVEC was increasingly hypothesized to be a driving factor of radicalism and terrorism in real life, tech companies were under pressure to take measures to manage the spread and impact of this content. Tech companies banded together to found NGOs and collaborative initiatives dedicated to handling TVEC to comply with new rules and to manage the explosion of TVEC on their platforms. These included:

- *The EU Internet Forum* (2015) and, later, its *EU Crisis Protocol* (2019): the Forum was created in response to terrorist attacks in Paris, Copenhagen, and Brussels, where TVEC was found to be a driving force of the perpetrators' radicalization and resort to violence.[16]
- *The Global Internet Forum to Counter Terrorism (GIFCT)*: Established in 2017 by Twitter, Microsoft, Facebook (Meta), and YouTube, this was a means by which industry could cooperate to prevent the spread of TVEC. It was also likely a mechanism through which members could cooperate to adequately respond to the new EU legal regime based on "The European Agenda on Security."[17] It was invigorated after the Christchurch shooting, initiating a "Crisis Incident Protocol" to respond to deluges of TVEC content in the immediate aftermath of terrorist incidents. It has become one of the major tools that member companies use to handle TVEC in the immediate wake

[15] "Informal Meeting of the Heads of State or Government Brussels, 12 February 2015 - Statement by the Members of the European Council"; "European Union Internet Forum (EUIF)"; "EU Internet Forum"; "EU Internet Forum"; COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS The European Agenda on Security.

[16] "European Union Internet Forum (EUIF)."

[17] Directive (EU) 2017/541 of the European Parliament and of the Council of 15 March 2017 on combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA; "The Online Regulation Series | The European Union - Tech Against Terrorism."

of a crisis today.[18] In addition to its founders, members of the GIFCT are Google, WhatsApp, Instagram, LinkedIn, Amazon, Zoom, Tumblr, Discord, WordPress, GIPHY, Clubhouse, Discord, Mailchimp, Airbnb, JustPaste.it, and MEGA.[19]

- *The Christchurch Call to Eliminate Terrorist and Violent Extremist Content Online*: a non-binding normative multilateral agreement signed by 58 countries and 14 technology companies; it outlines a set of principles to which signatories agree to abide. This initiative, while non-binding, set in train a series of real mitigation measures that many large tech companies adopted and use to this day to counter TVEC.[20]

While industry-government cooperation appears to have progressed, little has been written on how effective efforts to manage TVEC have been over the last four years and where challenges and gaps remain. There is a lack of clarity around how tech companies define TVEC, what technical tools are used for moderation and how effective they are, and where there are opportunities for improvement. Meanwhile, TVEC continues to have very real and problematic impacts on society.

[18] Radsch, "GIFCT."

[19] "Membership."

[20] Call, "Christchurch Call Text."

2. Problems with Taxonomy: Defining TVEC

Technology companies are not casting a wide net when it comes to classifying TVEC. Efforts to manage TVEC thus hit a roadblock from the outset: there is no consensus on how to define TVEC, and most companies' definitions are surprisingly narrow. Where there are individual definitions, these tend to have a heavy emphasis on notorious Islamic extremist groups, like ISIS or Al Qaeda, rather than, say, white supremacist terrorists. There is some variation among large technology platforms, from Facebook having a relatively precise definition, to YouTube which lacks a definition altogether.[21]

The United Nations Security Council's Consolidated Sanctions List of terrorist entities is generally recognized as the authoritative list to follow (if not this list, then other national or international lists of terrorist entities). There is an extremely high threshold for being placed on this list and is dependent on past terrorist behavior. (It is also worth bearing in mind that there is no international consensus on how to define 'terrorism'.)[22] Its semblance of authority on taxonomy has led to an ossification of the TVEC definition to mean pre-identified Islamic terrorist groups.[23]

It is perplexing that any technology company would think that this definition suffices for managing the broad range of TVEC today. The definitional focus on pre-identified *entities*, not terrorist or violent extremist *actions*, means that unidentified, unaffiliated, and lone-wolf actors are not covered by TVEC definitions until the terrorist attack has been carried out. That is, much too late to be of use for the would-be attacker. (Notwithstanding the importance of managing post-attack material that could serve to radicalize other would-be attackers.) In other words, if TVEC is defined as content associated with an organized terrorist group, the amount of data points is much fewer than if the definition is more inclusive, such as including specific actions or words.

A critical part of successfully dealing with TVEC is through data collection and

[21] OECD, "Transparency Reporting on Terrorist and Violent Extremist Content Online"; GIFCT, "Broadening the GIFCT Hash-Sharing Database Taxonomy: An Assessment and Recommended Next Steps."

[22] "Terrorism."

[23] Saltman, "Introducing 2022 GIFCT Working Group Outputs"; OECD, "Transparency Reporting on Terrorist and Violent Extremist Content Online."

accurate measurement; that is, understanding empirically how successful we are at removing or deprioritizing TVEC. The narrow definition makes it impossible to accurately measure how well technology companies are dealing with TVEC empirically. For instance, most large platforms self-report that they automatically remove around 95%+ of TVEC.[24] Beyond wondering what happened to the leftover 5% (which can represent hundreds of thousands of pieces of content per platform), we must also consider just how many millions of posts were not targets for detection to begin with under a narrow TVEC definition, but which could reasonably be categorized as TVEC.[25] For reference, Facebook stated that it removed over 33 million pieces of TVEC from its platform in 2020 – a massive amount even using its narrow definition.[26] Cross-company comparisons also become meaningless if we cannot control for definitional differences.

Even the Global Internet Forum to Counter Terrorism (GIFCT) - the major cooperative initiative among large tech platforms to remove TVEC - uses the UN List to guide its hash-sharing database. It recently expanded its definition to include content relating to several recent hate-fuelled terrorist attacks, including those in Christchurch, Glendale, and Halle. Again, in 2022, it expanded its taxonomy to include terrorist manifestos and some other PDF and text-based materials. Until recently, most companies did not define terrorist manifestos and other radicalizing text based and PDF materials as TVEC, only graphic content of terrorist incidents themselves. Small changes like these can have an outsized effect on the online information landscape. The GIFCT's narrowly-framed database only provides after-the-fact damage control; it addresses already-known symptoms, rather than proactively removing new content."[27]

Using narrow definitions, misogyny-based violent extremist content featured in incel ("involuntarily celibate") circles is often ignored, because incels are usually not designated a violent extremist or terrorist group, even though incel content has inspired offline terrorist attacks against women and couples. In fact, only 0.1% of GIFCT's hashes relate to incel violence (and that is only because it was linked to a "terrorist incident" – a shooting in Glendale, Arizona, which targeted women

[24] Tech Against Terrorism, "Gap Analysis and Recommendations for Deploying Technical Solutions to Tackle the Terrorist Use of the Internet."

[25] "Gap Analysis and Recommendations for Deploying Technical Solutions to Tackle the Terrorist Use of the Internet."

[26] "Gap Analysis and Recommendations for Deploying Technical Solutions to Tackle the Terrorist Use of the Internet."

[27] GIFCT, "Broadening the GIFCT Hash-Sharing Database Taxonomy: An Assessment and Recommended Next Steps"; "2022 GIFCT Transparency Report," December 2022.

and couples), and more than 90% were linked to the UN designated entities list as of mid-2021.[28] The GIFCT recently took the important step of labelling hashes by ideology, however, only 0.05% of hashes so far have been given ideology labels. Of those, 96.56% are labelled as Islamic Extremism, though the GIFCT has qualified that it expects this number to change as it reviews its taxonomy broadens.[29]

Discussions on taxonomy quickly turn into political and legal debates (similar to efforts to define 'hate speech'), and it is understandable why most groups have opted for a conservative 'least-common-denominator' definition. Tech companies do not necessarily *want* a narrow definition; instead, their incentive structures are misaligned. Even when tech companies want to adopt a broader definition, a mandated definition may be less of a headache to manage than one that is voluntarily proposed. This is because many companies are under constant legal pressure from shareholders to maximize shareholder value. There is a protracted legal debate about whether or not this responsibility is legitimate or a myth in legal terms, but that does not stop lawsuits from rolling in. In voluntarily adopting an expanded definition of TVEC – or generally doing more to fulfil corporate *social* responsibility or uphold ESG principles – firms' commitment to this supposed legal obligation may be questioned.[30]

Beyond small academic and industry circles, companies' very narrow definitions of TVEC, which almost exclusively focus on 'Islamic extremism,' appears to be a largely unknown phenomenon and is generally missing from the broader debate around countering TVEC online.

In some regions there has been a more rigorous debate about defining certain kinds of 'harmful' content than what exists in the United States. For instance, in places like Israel and in much of Europe, antisemitism and neo-Nazi content is more strictly regulated, if not outright banned.[31] These countries would probably have an easier time in formulating a more inclusive definition of TVEC and in enforcing a legal definitional requirement on companies. The United States' elevation of the First Amendment above these sorts of considerations means that all sorts of abhorrent

[28] GIFCT, "Broadening the GIFCT Hash-Sharing Database Taxonomy: An Assessment and Recommended Next Steps."

[29] "2022 GIFCT Transparency Report," December 2022.

[30] See, for instance, "Social Responsibility and Enlightened Shareholder Primacy"; "Corporations Don't Have to Maximize Profits."

[31] Goldsmith and Wu, Who Controls the Internet?

terrorist and violent extremist material can generally withstand judicial scrutiny.[32] However, even the U.S. Justice Department supports amending Section 230 of the Communications Decency Act of 1996 – the subject of copious controversy, and which has given a blanket pass to technology companies limiting liability for content on their platforms – to include more explicit language banning "unlawful" content and content that "promotes terrorism."[33]

The question of whether companies are targeting the full scope of the content that matters remains open. Some companies would probably argue that they deal with content not included in the TVEC definition in other ways, for instance, treating a terrorist manifesto as 'hate speech' or a video of a terrorist attack as 'graphic content,' violating terms of service but excluding that type of content from a more fulsome TVEC definition. For instance, Meta has banned "misinformation that has the potential to contribute to imminent violence or physical harm."[34] This definitional bifurcation carries risks; for instance, mitigation tactics against 'lower-tier' content are usually less stringent. This becomes problematic when the definition is so narrow. Perhaps more problematic is the issue of company inaction: some companies simply fail to uphold their terms of service and remove content that is found to violate their standards, whether it be because of lack of intention or capability.

Finding a balance is important: adopting a definition that is too broad may lead to a tendency to treat TVEC less robustly. Clearer definitions of second- and third-tier types of harmful content, like 'incitement to violence' and 'hate speech' may give greater clarity to what constitutes TVEC. Corporate introspection to assess biases, such as examining whether incel and white supremacist terrorism is categorically treated the same as Islamic terrorism (rather than, say, being unfairly treated as a lower-tier category of harmful speech) would be a helpful complement to any policy solution.

There will inevitably be grey areas. For instance, alt-right memes that propagate racist conspiracy theories are often a point of contention, with propagators arguing that such content constitutes "satire," often giving it the protection of plausible deniability.

[32] Goldsmith and Wu.

[33] "Section 230 — Nurturing Innovation or Fostering Unaccountability?"

[34] "Understanding Social Media and Conflict."

A more thorough examination of what kinds of content are omitted from companies' definitions, whether gaining consensus on a more inclusive definition should be considered, and how definitions could be implemented or enforced in practice, are important avenues for future research.

3. Alternatives to a Legal Definitional Requirement

Legal definitional requirements can hold companies accountable, but they can also create false, lower standards than what society might expect of technology companies. Enforcement is often challenging and seen as moral disapprobation by critics. To be effective, laws must be precise enough to be enforceable. One possible alternative is a normative regime whereby technology companies have a broader, ethical obligation that goes beyond what may be required by law.

Encouragingly, some companies, recognizing that legal change can be a long and arduous process and acknowledging the harm caused by TVEC, have adjusted their company policies to be more expansive than the law requires. For instance, after the Christchurch shooting, Microsoft played an important role as a "norm entrepreneur": taking actions then seen as radical to encourage other technology companies to act against TVEC and collaborate on multilateral commitments to eliminate material relating to the Christchurch shooting.[35] The GIFCT, as a powerful normative actor (most of the largest U.S. tech companies sign on to this initiative), could pave the way for broader adoption of an inclusive definition of TVEC if it were to continue driving forward with evidence-based taxonomic changes. Since 2021, several mainstream platforms have taken steps towards adopting a more inclusive approach and defining TVEC and something broader than simply ISIS content, but progress across the board has been slow and is far from producing a uniform definition.[36]

Civil society, governments, companies, and users alike can play important norm-changing roles by going beyond what is required to 'do the right thing,' much like the state of California goes beyond federal environmental regulations despite having no legal obligation to do so.[37] By framing the issue in terms of a social contract with citizens, who deserve to be free from harm, change may be possible via normative mechanism.[38]

[35] Smith and Browne, Tools and Weapons; Lohr, "How Top-Valued Microsoft Has Avoided the Big Tech Backlash."

[36] OECD (2022), "Transparency reporting on terrorist and violent extremist content online 2022", OECD Digital Economy Papers, No. 334, OECD Publishing, Paris, https://doi.org/10.1787/a1621fc3-en.

[37] Schmidt, "ENVIRONMENT"; Tabuchi, "U.S. Climate Change Policy."

[38] Finnemore and Sikkink, "International Norm Dynamics and Political Change."

Popularization of the ‘ethical AI’ concept with the advent of products like ChatGPT is an example of how users can demand robust standards from technology companies where they would otherwise be free to adhere to minimum standards set by law. Today, when AI products get something wrong or start spouting violent or ‘hateful’ comments, users are in uproar, and companies rush to find fixes.[39]

Beyond company-level policies, engineers and technologists working within technology companies are often at significant liberty to design products and policies in ways that reflect their values. As engineers and technologists become increasingly aware of the concepts of ethical AI and responsible innovation, individual designers can make an impact on how technology companies operate on the inside. In this same vein, individuals working within the tech industry on issues relating to TVEC could take on more personal responsibility in how they manage and design products from a trust and safety perspective and provide downward leadership to their team as they move into roles with greater authority. Educators could play an important role towards this end: by teaching future technologists in universities about the impacts of TVEC and how to design products in ways to mitigate harm, a new generation of technologists could shape norms relating to ethical tech design and moderation. Perhaps a kind of Hippocratic oath for technologists is in order.[40]

Of course, there are always actors that try to exploit norms in the absence of legal requirements, and some companies remain committed to being the so-called “free speech arm of the Free Speech Party.”[41] Users that seek to read or post TVEC have increasingly turned to these sites as safe havens from regulators and platforms with stricter content policies and removal abilities. For instance, platforms like Gab, Parlor, 4chan and 8chan are infamous for hosting white supremacist TVEC and having lax content moderation policies. There are also small companies that simply lack the resources to meaningfully tackle the problem.[42]

All this is not to say that technology companies currently do not want to do the right thing. In fact, normative shift may be a viable option today *because* there is a

[39] “How OpenAI Is Trying to Make ChatGPT Safer and Less Biased.”

[40] This nomenclature was first developed by Abbas, Senges, and Howard, “A Hippocratic Oath for Technologists.”

[41] Halliday, “Twitter’s Tony Wang.”

[42] Tech Against Terrorism, “Gap Analysis and Recommendations for Deploying Technical Solutions to Tackle the Terrorist Use of the Internet.”

greater awareness among technologists of the harms caused by TVEC. One of the main challenges technology companies face today is not a lack of willingness to do something about the issue but that finding the tools to do so successfully is very challenging.

4. Technical Tools: No Panacea, but Heading in the Right Direction

There is often an assumption among policymakers that technology companies – if compelled to do so – can get their engineers to wave their hands and create the perfect technical solution to content moderation problems.[43] Unfortunately, this is not the case. Even if there was unanimous agreement across governments and the tech industry to the legal and sociological questions on TVEC, and the best of intentions among all actors, there still would not be a technical panacea to the problem. This is a second major roadblock to managing TVEC.

The tools commonly used today by major technology companies each come with their pitfalls and trade-offs across efficiency, cost, scalability, and accuracy. Current tools are also not uniform in their application and purpose: some try to address the symptoms of TVEC; others prevent its occurrence in the first place. Most deal with a subset of the problem, like identification, and must be used in combination with other tools for removal or other method of management. Most demand some level of human interaction.

Like with spam and child sexual abuse material, there will unfortunately always be content that slips through the cracks.[44] The goal thus becomes achieving a result that maximizes success across the aforementioned variables according to interests. This requires determining a balance of interests: this might involve settling on a tolerable false positive rate that also allows for maximization of other variables such as speed and scalability.

In general, a good tool optimizes for the following: resilience (it is not easily evaded or undermined), accuracy (it correctly targets the problematic content and has a low false positive/negative rate), scalability and speed (the technology keeps up with the submission rate and covers close to 100% of problematic content and across the entire tech ecosystem), and ease of implementation. There is little research covering the efficacy of major tools currently used by the largest technology platforms, and the following analysis highlights the major benefits and pitfalls of each.

[43] Gorwa, Binns, and Katzenbach, "Algorithmic Content Moderation"; "Challenges in Combating Terrorism and Extremism Online."

[44] "The US Now Hosts More Child Sexual Abuse Material Online than Any Other Country."

5. Assessment of Technical Tools

5.1 Hash-matching

Hash-matching is one of the most commonly deployed technical tools for identifying TVEC among large U.S.-based technology giants. A *hash* is a unique identifier or "digital fingerprint" that is issued to a piece of media – an image, a video, an audio file, and so on. Hashing takes an arbitrary set of bits and transforms it into a smaller fixed-length value that is unique to those bits. Its small size enables it to be compared with large numbers of other hashes. Hash-matching is the comparison and identification of identical (or near-identical) hashes across platforms for efficient detection of all instances of that content.[45]

Cryptographic hashing can locate identical matches but cannot match a piece of content that has been altered in even the slightest form. Changing the value of just one pixel in an image or adding one extra space in a written document will result in a completely different hash. Because actors wanting to spread TVEC are practiced at manipulating content to avoid detection, this form of hashing is very easily undermined by tactics such as adding a watermark or cutting off the corner of the frame.[46]

Perceptual hashing (a form of 'fuzzy hashing') is broadly used as the preferred hashing technique by major U.S. technology companies because it is better at overcoming cryptographic hashes' resilience issue, though it is far from perfect. Perceptual hashing identifies near-identical matches, including slightly altered media. It creates a hash using 'perceptual' features, like a rhythm in an audio file, or a corner of an image. It might match to images or audio clips that have 98% or 99% likeness.[47] The GIFCT also uses *locality-sensitive hashing*, which finds 'nearest neighbor' hashes. In other words, it groups content into data clusters and then

[45] Gorwa, Binns, and Katzenbach, "Algorithmic Content Moderation"; "An Overview of Perceptual Hashing | Journal of Online Trust and Safety"; "Overview of Perceptual Hashing Technology."

[46] Gorwa, Binns, and Katzenbach, "Algorithmic Content Moderation"; "Overview of Perceptual Hashing Technology"; "An Overview of Perceptual Hashing | Journal of Online Trust and Safety."

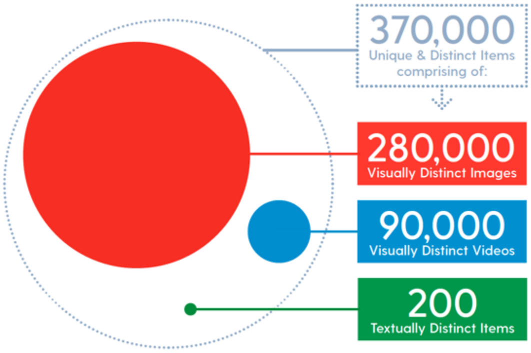
[47] Gorwa, Binns, and Katzenbach, "Algorithmic Content Moderation"; "Overview of Perceptual Hashing Technology"; "An Overview of Perceptual Hashing | Journal of Online Trust and Safety."

locates the closest hash from an original hash.[48]

Perceptual hashing and locality-sensitive hashing, while more resilient than cryptographic hashing, are still quite easily evaded by adversarial actors, research suggests. Hash matching is a cat-and-mouse game, whereby some players constantly figure out ways to avoid detection. However, detection technologies continue to develop and improve to overcome adversarial actors.[49]

The most ambitious experiment in using hash-matching for TVEC identification is probably the GIFCT's hash-sharing database, which collects and shares known TVEC hashes between large technology firms. The GIFCT has a "Content Incident Protocol" (CIP), whereby members act quickly to defend against a recent terrorist or violent extremist attack in real life by hash-matching and removing TVEC relating to the event. As of 2022, the GIFCT hash database contained around 2.1 million hashes, making up 370,000 "distinct items" of hashed content (see *Figure 1*). [50] The main algorithms used in this process are PDQ12 and PhotoDNA, though Meta in December released another called Hasher-Matcher-Actioner (HMA) which reportedly builds on its previous PDQ and TMK+PDQF algorithms.[51]

Figure 1. Hashed content in the GIFCT database (GIFCT, 2022).



From a purely technical standpoint, matching hashes to other hashes itself is a rapid process compared to other tools. In practice, however, hash matching

[48] "Introduction to Locality-Sensitive Hashing"; Tsai and Yang, "Locality Preserving Hashing"; GIFCT, "Advances in Hashing for Counterterrorism."

[49] See, for instance, Avril Wong, "Deep Perceptual Hashing Is Not Robust to Adversarial Detection Avoidance Attacks."

[50] "2022 GIFCT Transparency Report."

[51] GIFCT, "Broadening the GIFCT Hash-Sharing Database Taxonomy: An Assessment and Recommended Next Steps"; "Meta Launches New Content Moderation Tool as It Takes Chair of Counter-Terrorism NGO."

relies on hashing TVEC in the first place, which is usually a slow manual process performed by a human (or an automated process ratified by a human). It is important to understand that hash-matching is only helpful in identifying the re-posting of previously detected content, rather than new TVEC. Now, consider the immense scale of content uploaded to technology platforms every day – or for that matter, every *second*. For context, 400 hours of content is reportedly uploaded every minute to YouTube alone.[52] Given the pace at which TVEC is uploaded to technology platforms, even near-instantaneous processes would struggle to keep up, even without a "crisis incident." Without automated moderation systems, you would probably need the entire human population to be employed as content moderators to keep up with the scale of posted content.

Thus, there is a tension between algorithmic processes, which often lack the quality of judgment performed by a human but have the advantage of being scalable; and human moderation, which presents the reverse conundrum. Even human-automation synthesis can be too slow: usually an automated system will create a 'shortlist' of TVEC, with a human moderator tasked with double-checking accuracy. Even this process is far too slow to keep up with the continuous slew of uploaded TVEC.

Overall, research and case studies suggest that hash-matching technology tends to be highly accurate and results in few false positives.[53] False positives for the GIFCT only tend to occur because of human error: when the original hash flagged by a human as TVEC was found to not meet the GIFCT's or technology company's narrow definition of TVEC. Even the human error element in this regard is small: fewer than 0.1% of hashes were found to not meet the definition.[54]

Hash matching also has limited scalability in that not all hashes across all social media platforms or services can be matched at once, because of privacy and proprietary issues. For instance, Twitter cannot hash match across WhatsApp or Airbnb. Additionally, end-to-end encrypted messaging applications and cloud storage applications are more difficult to patrol than unencrypted and more open social media platforms. Hash-sharing between platforms may help to ameliorate the problem of cross-company differences, but design and organizational problems

[52] "YouTube Now Gets Over 400 Hours Of Content Uploaded Every Minute."

[53] See, for instance, "Case Study: Using the GIFCT Hash-Sharing Database on Small Tech Platforms."

[54] "2022 GIFCT Transparency Report," December 2022.

— especially the limited membership of the GIFCT — means that hashes in the database will not necessarily be identified or shared across a large portion of the information ecosystem.[55]

Finally, the GIFCT's hash-matching database has limitations from an organizational and enforcement perspective. Firstly, GIFCT cannot compel any company to act on TVEC hashes. Individual companies are ultimately responsible for addressing TVEC hashes on their platforms; the content, once identified, may be removed, deprioritized, sent for further review, or left online.[56] Secondly, the GIFCT's membership list has some notable omissions, especially non-U.S. technology companies, non-platform companies, and smaller technology companies.[57] Organizations must apply to become a member of GIFCT. Criteria for membership includes the following (paraphrased for brevity): organizations must be able to demonstrate that they have publicly available policies that explicitly ban terrorist and/or violent extremist activity, the ability to review and act on reports of TVEC, a desire to explore new technical solutions to the problem, regular and public data transparency reports, and a public commitment to respect human rights.[58] It is possible that other companies have applied, but failed, to gain admission. For instance, reporting in *The Hill* suggested that TikTok had applied for membership in 2019 but concerns relating to its data collection and censorship practices meant that its application was denied (though those claims are yet uncorroborated).[59] Smaller companies are less likely to have the resources to review and act on TVEC, which is a prerequisite for membership. Hash-matching is only useful if it can be employed at scale; and while tools are being developed to help smaller companies conduct their own hash matching, it is much harder for smaller firms with fewer resources, or for those that don't have access to the GIFCT database.[60] The damage can be substantial: footage of a shooting that targeted people of color in Buffalo, New York, was viewed 3 million times on Streamable before being taken down.[61]

[55] GIFCT, "Broadening the GIFCT Hash-Sharing Database Taxonomy: An Assessment and Recommended Next Steps"; Radsch, "GIFCT."

[56] Gorwa, Binns, and Katzenbach, "Algorithmic Content Moderation."

[57] OECD, "Transparency Reporting on Terrorist and Violent Extremist Content Online."

[58] Xx

[59] Birnbaum, "TikTok Seeks to Join Tech Fight against Online Terrorism."

[60] Tech Against Terrorism, "Gap Analysis and Recommendations for Deploying Technical Solutions to Tackle the Terrorist Use of the Internet."

[61] Harwell and Oremus, "Only 22 Saw the Buffalo Shooting Live. Millions Have Seen It Since."

5.2 Machine Learning (ML) Detection and Classification

Among the largest U.S. tech companies, trained machine learning (ML) algorithmic detection, which automatically screens and categorizes newly uploaded TVEC, is probably the most widely used tool to counter TVEC.[62] Automated content classifiers are trained to recognize TVEC by practicing on datasets, so that they can learn what is and isn't TVEC (hence the term "machine learning," which is a subset of artificial intelligence). A predictive score is then assigned to new content, with parameters set for deletion or other 'governance' action. Most large U.S. tech companies' ML tools use natural language processing (NLP) for prediction, which allows for greater contextual analysis. (NLP can be thought of as a way to help machines 'think' or process information more like humans.)[63] As will be explained, however, technology companies' ML detection tools continue to face contextual challenges.

The GIFCT Technical Working Group defines content classification as "automated detection of likely terrorist content based on prior similar content or inclusion of high-risk attributes such as terrorist logos, terminology, and imagery." It also describes content classifier tools as "…extremely complex and vulnerable to adversarial shift."[64] (Adversarial shift is when training data differs from what the tool sees in the real world and it loses accuracy.)[65] For instance, YouTube's ML terrorism detector was criticized for "erasing history" in taking down "witness videos" of the war in Syria, and not understanding the videos' context.[66] The GIFCT in 2022 put out a call for research proposals to develop a system to classify multimedia content as TVEC, indicating a need for this tool's improvement, as well as perhaps a belief in the potential for its future development and deployment.[67]

[62] Tech Against Terrorism, "Gap Analysis and Recommendations for Deploying Technical Solutions to Tackle the Terrorist Use of the Internet"; OECD, "Transparency Reporting on Terrorist and Violent Extremist Content Online"; 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."

[63] Gorwa, Binns, and Katzenbach, "Algorithmic Content Moderation."

[64] Tech Against Terrorism, "Gap Analysis and Recommendations for Deploying Technical Solutions to Tackle the Terrorist Use of the Internet."

[65] 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."

[66] Khatib and Kayyali, "Opinion | YouTube Is Erasing History"; "'Lost Memories.'"

[67] "GIFCT Working Groups Output 2022."

Facebook had difficulty using ML tools to detect TVEC livestreaming like the Christchurch shooting at first, because it didn't have enough data to train on relating to firearms; however, the U.S. and U.K. governments agreed to provide it with first-person firearms footage for training their ML tools.[68] The quality and scope of training data is critical to the success of ML tools, but there will always be some false positive rate. Teaching an ML tool to differentiate between, say, a first-person real-life shooting versus a first-person shooter video game, is an important differentiation but can be technically challenging. These are some of the major limitations of current tools for identifying TVEC, and most large technology companies stress that their ML tools need to be supplemented by human moderators.[69] As Dr. Erin Saltman, director of programming at the GIFCT, wrote in 2021, "We can't simply algorithm our way out of the problem."[70]

The reality is that companies use a mix of human-in-the-loop processes and fully automated processes for TVEC management, to varying degrees. Facebook automatically removes content that its ML classifier tools find to be a clear-cut case of TVEC, whereas a predictive score that is not so clear-cut might be flagged for human review.[71] Although ML tools are often criticized for failing to properly understand context, humans sometimes aren't much better than machines at choosing what is or is not TVEC. (In 2016, human moderators at Facebook took down images of the Pulitzer-Prize-winning Vietnam War photograph of "Napalm Girl," classifying it as child pornography, causing a public uproar.)[72] The line between moderation and censorship is not always clear; both machines and humans make mistakes in this regard.

ML prediction tools also face challenges relating to inequality and unevenness in application: like almost all AI tools, ML detection ends up being a reflection of our societies, and especially of those in power who design the training data.[73] This has resulted in uneven outcomes, like greater effectiveness of ML tools across English-language media, and a greater likelihood that TVEC content that is not Islamic extremist is not categorized as such. According to a 2020 Global Network

[68] UK Government, "Firearms Officers Begin Filming Training for Counter Terrorism Initiative"; "Facebook to Train AI Systems Using Police Firearms Training Videos."

[69] OECD, "Transparency Reporting on Terrorist and Violent Extremist Content Online."

[70] "Challenges in Combating Terrorism and Extremism Online."

[71] Gorwa, Binns, and Katzenbach, "Algorithmic Content Moderation."

[72] Shahani, "With 'Napalm Girl,' Facebook Humans (Not Algorithms) Struggle To Be Editor."

[73] 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."

on Extremism and Technology (GNET) report on AI and countering TVEC, many large technology platforms' ML tools omit or do not adequately understand content that is written or spoken in 'minority' languages.[74] Additionally, ML tools are often trained on datasets of a limited nature – for instance, in the United States or United Kingdom context – and fail to understand different cultural context and flag content that incited ethnic violence (as was the case in Colombo, Sri Lanka, prior to the 2019 Easter bombings). The report states, "At the moment only 14 out of Europe's 26 official languages are covered in Facebook's fact-checking language repertoire."[75] Developments in natural language processing are helping to overcome the aforementioned translation and contextual issues, such as *masking*, but there is no perfect product.[76]

ML detection has the benefit of being immediate and scalable (at least for large technology companies). Its major downside is its inability to understand context, especially in non-text-based media, resulting in less-than-desirable performance in resilience and accuracy.[77] It is difficult to make any broad-based empirical judgements on the rate of false positives, as the data for each company is very opaque and different tools are used to target different kinds of content. However, users tend to perceive that automated ML detection tools have a high false-positive rate, being influenced by high-profile detection mistakes and because society holds machines to higher standards than their human counterparts.[78] Humans - especially in countries like the United States - are taught that censorship is anathema to basic liberty; making mistakes like false positives seems particularly tyrannical. This societal pressure leads to downward pressure on technology companies' thresholds for moderation.[79]

ML detection tools benefit from not requiring collaboration across technology companies in order to be effective. Whereas the GIFCT is made up of almost exclusively U.S. companies, and membership greatly increases the value and efficacy of hash matching, ML prediction tools do not require the same level of cooperation to be effective. This means that for at least thirteen of the top 50

[74] "Artificial Intelligence and Countering Violent Extremism: A Primer."

[75] "Artificial Intelligence and Countering Violent Extremism: A Primer."

[76] "Artificial Intelligence and Countering Violent Extremism: A Primer."

[77] "Artificial Intelligence and Countering Violent Extremism: A Primer."

[78] Naughton, "To Err Is Human – Is That Why We Fear Machines That Can Be Made to Err Less?"; Günther and Kasirzadeh, "Algorithmic and Human Decision Making."

[79] "Moderating Online Content"; Nadeem, "Most Americans Think Social Media Sites Censor Political Viewpoints"; Llansó, "No Amount of 'AI' in Content Moderation Will Solve Filtering's Prior-Restraint Problem."

online content-sharing services, which are Chinese technology giants, ML prediction will likely be more effective than hash sharing given current organizational arrangements.[80] However, smaller platforms find themselves at a disadvantage: they are less able to reap the benefits of ML tools, given the high upfront cost and ongoing capabilities required to manage ML classifier tools, and thus deploy ML tools less frequently than tech giants.[81]

5.3 **Recommendation System Adjustments**

Recommendation systems are a powerful tool of curation, heavily influencing how humans interact with online information today. What we see online is determined by how these systems are designed. Recommendation algorithms tend to be designed to maximize engagement, with often problematic societal implications. This means recommending or prioritizing content that the algorithm predicts the user wants to see, which may be radicalizing and extreme content. Some platforms have taken measures to mitigate against this threat following public criticism that platforms create "rabbit holes" and "echo chambers."

Search and recommendation algorithms are usually the subject of criticism rather than opportunity in the academic literature about TVEC. The reality is that they have both the capability to be a weapon (e.g., reinforcing or amplifying TVEC) and a tool (e.g., deprioritizing TVEC or off-ramping from radicalizing content). Policymakers often jump to regulate technology companies' recommendation algorithms without understanding the implications or core concepts underlying the technology. This section explores some of these knowledge gaps.

Search algorithms, such as those used for Google's or Twitter's Search function, respond to explicit queries, such as a keyword search, to retrieve specific information. A search algorithm may produce or rank results based on previous user activity. *Recommendation algorithms*, like those used for personalized ads, Facebook's Newsfeed or TikTok's 'next up' function, suggest content based on users' previous activity and predicted preferences based on the data collected about the user or group of users.

[80] OECD, "Transparency Reporting on Terrorist and Violent Extremist Content Online."

[81] Tech Against Terrorism, "Gap Analysis and Recommendations for Deploying Technical Solutions to Tackle the Terrorist Use of the Internet."

Machine learning algorithms are subject to the many shortcomings in ML tools already detailed above. Beyond these challenges, there is a growing body of literature on the ways in which search and recommendation algorithms may amplify or reinforce TVEC and radicalize users.

There is broad consensus in the academic literature that large technology platforms tend to have a profit incentive to maximize engagement, usership and 'clicks'. In other words, technology companies generate more revenue by recommending content to users that will keep them using the product. Since humans tend to be drawn to more radical and attention-grabbing content, recommendation algorithms tend to be designed to promote this content by giving users content based on what they have previously seen or searched for, or according to certain other data collected on the user. No consensus on a causal link between recommendation systems and radicalization has been reached, though this is widely assumed, and most qualitative studies hypothesize a causal connection — that recommendation algorithms on their own will guide users towards extremist content.

We do know that investigations into recent lone-wolf terrorists have found that their online activities contributed to their radicalization. For instance, New Zealand's Royal Commission of Inquiry into the Christchurch Shooting found that YouTube had been a "significant source of information and inspiration" for the shooter. The shooter's manifesto was also littered with references to online extreme-right-wing "in-jokes" from both technology platforms and video gaming, demonstrating the influence that extreme-right-wing online subcultures had on him. The Royal Commission found that the shooters "exposure to such content may have contributed to his actions on 15 March 2019 - indeed, it is plausible to conclude that it did."[82]

However, as of this writing, an empirical causal link between the recommendation algorithms used by major technology platforms and *amplification* (a tendency to recommend more and more extreme content) is yet to be proven in the literature. The few empirical (yet outcome-based) studies performed have been focused mostly on YouTube, which has a relatively open API, and on English-language media. These studies overwhelmingly suggest that there is an amplification

[82] "Assessment of the Individual and the Terrorist Attack"; "General Life in New Zealand."

effect.[83] One empirical study recently found mixed results: that YouTube disproportionately amplified extreme content, and Reddit and Gab did not.[84] This study also relied on outcome-based results rather than looking at the design of the recommendation systems themselves. The problem is simply that there is limited information available to scholars and to the public to adequately answer this question, and results are likely to vary by company.

Putting aside the question of amplification, the literature finds that recommendation algorithms do tend to create "filter bubbles" or "echo chambers" around users that give them more content adhering to a certain ideology, if that is what the user's preference is judged to be. That is, recommendation algorithms are more likely to give users content similar or relevant to material that they have looked at or searched for previously (as opposed to offering *more* extreme material as the default). This means that, for a user searching for violent extremist content, an algorithm designed to maximize engagement may recommend similar content.

So, why is it so hard to understand the relationship between recommender systems and TVEC? Barriers arise at the most basic level: it can be extremely challenging to understand how recommendation algorithms make decisions. This is sometimes referred to as the algorithmic "black box" phenomenon. Additionally, recommendation algorithms are the golden goose of technology platforms and tend to be harbored with care and strict confidentiality. This opacity means that researchers are left to make sense of algorithmic outputs, without visibility over the inputs or the decision-making process.[85]

The literature on the topic is also heavily skewed towards what information is available in the public domain, focusing on the few companies that have a more open API. There is little question as to why Paul Covington and co-authors' paper on "Deep Neural Networks for YouTube Recommendations" has for years been one of the most highly referenced works on the recommendation system of a major technology platform: it offers a rare insight into the logic and design of a major recommendation system, where there is scant insight elsewhere.[86]

[83] For a comprehensive overview of the literature to date, see: Whittaker, "Recommendation Algorithms and Extremist Content: A Review of Empirical Evidence."

[84] Whittaker, "Recommendation Systems and Extremism"; Whittaker et al., "Recommender Systems and the Amplification of Extremist Content"; "Content-Sharing Algorithms, Processes, and Positive Interventions Working Group."

[85] "Content-Sharing Algorithms, Processes, and Positive Interventions Working Group."

[86] Covington, Adams, and Sargin, "Deep Neural Networks for YouTube Recommendations."

Technology companies themselves and their engineers also sometimes don't have complete understanding of the decision-making process: the advent of algorithms that use deep learning or neural networks to make decisions means that humans sometimes don't know how decisions are made. Even when algorithms are more rudimentary, they operate in ways that make it hard to draw a direct link between recommendation algorithms and amplification of extreme content. This is because recommendation algorithms tend to be based on a user's individual history and perceived preferences, making it hard to control what content is recommended to each individual user. The same algorithm that suggests a new pasta recipe for one user may be the same algorithm that suggests a video promoting a white supremacist conspiracy theory to another.[87] The cause-and-effect question is still at large. Are users, radicalized offline, responsible for driving the technology towards recommending TVEC? Or, is it a function of technological determinism; that is, is the technology driving users who wouldn't normally seek out TVEC towards that content?[88] Both are possible, with conceivable bidirectional feedback loops.

The first problem related to using adjustments to recommendation algorithms is this: we don't have adequate insight into how they are designed in the first place or the nature of their links to amplification of TVEC, and therefore, it is hard to say how best they can be adjusted to limit the spread of TVEC. But even if it is not the algorithms' 'fault' per se that they play a role in amplifying or reinforcing TVEC – and even if recommendation algorithms are only a reflection of the population or groups of users – one could still argue that technology companies should have an ethical – if not legal – obligation to steer users away from TVEC and borderline material.

Opacity of tech companies, coupled with media reporting on the purported radicalizing effects of social media curation, has prompted a wave of political support for "algorithmic transparency" policies – compelling firms to give insight into how their recommendation systems filter content, or even compelling them to hand over their algorithms for review by authorities.

Algorithmic transparency is often touted as a silver bullet to manage algorithmic recommendation of TVEC, though it has its challenges. Most fundamentally, there is no clear understanding of what it means for a ML algorithm to be "interpretable," even among engineers.

[87] The YouTube Team, "Continuing Our Work to Improve Recommendations on YouTube"; Jones, "What Is the 'great Replacement' and How Is It Tied to the Buffalo Shooting Suspect?"

[88] Whittaker, "Recommendation Algorithms and Extremist Content: A Review of Empirical Evidence."

The following hypothetical example may help to explain why this is the case. Say, you want a recommendation algorithm that *de-prioritizes* TVEC. You train it on a set of data – some ISIS videos, mass shooting livestreams, as well as some non-TVEC – to teach it how to predict TVEC. The decision – "is this TVEC?" – depends on a set of factors identified by algorithm, such as presence of ISIS propaganda or white supremacist phrases, or extent of gore or violence. Each factor is assigned a weighting, and together will form a large stack of equations – maybe millions or billions of equations – that are calculated and then form a prediction score. Using trial and error, the algorithms find the weightings that achieve the best result. A human will clearly struggle to understand why the algorithm makes decisions and what patterns it recognizes.

Sometimes an algorithm will produce surprising and problematic results. Rather than identifying TVEC based on attributes that humans might choose, the algorithm may pick up that the training data labelled "TVEC" tends to contain an attribute that should not be associated with TVEC. For instance, the training data may, as a result of human bias and poor training data, overwhelmingly include TVEC-labelled data that includes men with beards. The algorithm may then start to de-prioritize content that contains men with beards, predicting that it signals TVEC. Or, the algorithm may pick up on some other benign similarity in the TVEC data that is not representative of the total population of TVEC.

Not only can there be harmful impacts that arise from such pattern recognition, as this example shows, but it can also be hard for humans to pick out on what exactly the algorithm is basing its decisions. Once trained and applied to real data, the algorithm will adapt and adjust on its own. This can lead to other problems and changes to the algorithm that impacts its performance, such as "overfitting", which is when the model is trained too closely on the training data and is not well suited to the entire population of real-life data. The algorithm is not a finite thing that can be written on a chalkboard, but is instead a dynamic, complex process over which humans cannot have complete visibility.

The following example illustrates some of the real-life challenges associated with algorithmic transparency: in one of the most ambitious cases of regulation, the Cyberspace Administration of China (CAC) in 2022 enforced far-reaching and strict new regulations on technology platforms' recommendation algorithms, including an "algorithm registry" that requires firms to hand over details of their recommendation algorithms to CAC officials (for instance, information on the data used to train models). The registry itself is extremely opaque, and we do not know what exactly companies have been compelled to hand over to CAC officials, including whether authorities have requested direct access to the algorithms.[89] The rules also prohibit algorithms that are deemed to "endanger national security" or "violate public order," and mandate that algorithms "actively disseminate positive energy" (translated from the original language into English).

This case revealed striking shortcomings in efforts to regulate – and to simply understand – recommendation algorithms through algorithmic transparency policies. Media reporting alleges that ByteDance employees had to speak in very broad and vague terms when explaining their algorithms to officials who clearly did not know how to engage with the technologists. Officials had no idea what to look for, did not understand how ByteDance's recommendation systems worked, could not adequately interrogate the presented information, and were unable to delve deeper than surface level metaphors and simple explanations.[90] Moreover, ByteDance engineers themselves would not have been able to explain everything, even if officials were well versed in recommendation systems. Other countries and regions have implemented their own version of algorithmic transparency measures, with similar problems.[91]

There are other challenges associated with algorithmic transparency. Companies are opaque for reasons beyond profit maximization and IP protection: for instance, opacity reduces the threat of adversarial actors gaming algorithmic recommendation systems. Giving 'bad actors' access to the methods, techniques, and sources used by companies to filter content is ammunition for exploiting the algorithmic design to spread certain types of content – for instance, knowing how a search algorithm prioritizes (or deprioritizes) certain types of content enables an actor to get around those rules by adjusting their content to increase its ranking.

Another challenge in using algorithmic adjustments is the fact that there are many different types of recommendation algorithms, from collaborative or content-

[89] Du, "What China's Algorithm Registry Reveals about AI Governance"; "China Passes Sweeping Recommendation Algorithm Regulations."

[90] Hao, "China May Be Chasing Impossible Dream by Trying to Harness Internet Algorithms"; Du, "What China's Algorithm Registry Reveals about AI Governance."

[91] Rowa, "The Contextuality of Lone Wolf Algorithms: An Examination of (Non)Violent Extremism in the Cyber-Physical Space."

based filtering to deep learning algorithms, each with their advantages and pitfalls. There is not a cookie-cutter adjustment that can be applied to all recommender systems across all applications and Web sites.

The bottom line is that algorithms tend to reinforce – and may amplify – content that the algorithm predicts that an individual wants to see, including potentially TVEC. So, how have companies adjusted their practices to mitigate this threat?

Companies have generally made adjustments to their recommendation systems that can be boiled down into two categories: "de-prioritization"/"de-platforming" and "positive interventions."

De-prioritization and de-platforming: Recommendation algorithms can be designed to "deprioritize" or "de-platform" TVEC, by, for instance, lowering TVEC's ranking in search functions or removing it from search results. There are crossovers with the "ML tools" section described above: using ML to identify and take some action on TVEC.

Positive interventions: Recommendation and search algorithms can be programmed to stage positive interventions, for instance, by promoting de-radicalizing content for at-risk users, also known as "off-ramping." This technique is often used by technology companies and media outlets in response to users who search for suicide-related content: the search may produce a hotline for mental health crisis assistance or pages relating to self-help to steer away from a page on self-harm. Similarly, users who search for, say, a white supremacist manifesto, may be steered instead towards a page debunking racist conspiracies contained in the document or something similar.

The GIFCT's 2022 Working Group outcome paper on "Recommendation Algorithms and Extremist Content: A Review of Empirical Evidence" provides a helpful scan of major technology companies and the (public) changes that they have made to their recommendation algorithms, including the following examples:[92]

In 2015 Reddit announced that it would "quarantine" subreddits that are deemed to "be extremely offensive or upsetting to the average Redditor," restricting them to users who opt in, creating a kind of private chat room for users who want to see

[92] Whittaker, "Recommendation Algorithms and Extremist Content: A Review of Empirical Evidence."

this content.[93] Quarantined pages reportedly generate no revenue and lose their subscribers.[94] It is otherwise unavailable to users who are not subscribers and is removed from searches and recommendations.[95] No information is available on the extent to which this impacted TVEC on Reddit, though subreddits tend to have a reputation for containing content that is at the very least "borderline." Reddit will sometimes ban subreddits, though it can take a long time for bans to be imposed, even if the forum violates community guidelines – such as the notorious "r/beatingwomen" subreddit, which was a forum for posting content of exactly what the name suggests.[96] For instance, the quarantined subreddit "r/Chodi" was banned in March 2022 after hitting 90,000+ members with hundreds of posts every day; posted content called for violence against and genocide of Muslims. (Reddit also faced heavy criticism for failing to deal with the coded language used by subscribers to get around NLP identification tools.)[97]

This type of de-platforming and de-prioritizing via "quarantine" may be helpful in shielding the everyday user of Reddit from TVEC, but it siphons off at-risk users into echo chambers on other platforms. This trend is growing and is hard to avoid. For instance, when "r/Chodi" was banned from Reddit, its members simply migrated their private cesspool of TVEC to Telegram.[98] This is part of a broader trend and backlash against efforts to combat TVEC: in general, users seeking to post or view TVEC online have gravitated towards smaller platforms with more lax moderation protocols or towards different forms of media that are outside of the mainstream, such as file-sharing services and private message boards.

In 2019, YouTube announced that it would begin to deprioritize and reduce recommendations of TVEC and "borderline" content – content that comes close to but does not violate YouTube's community guidelines, like misinformation about historic events like 9/11 or a "phony miracle cure." It did not go so far as to start removing this content, but simply to rank it lower in search results and measures to that effect. According to YouTube's CEO, Susan Wojcicki, the move reduced watch time of that content by 50%. It reportedly uses a "combination of machine learning and real people." YouTube also reported that it "raises" authoritative

[93] Sottek, "Reddit Bans Several of Its Most Racist Communities."

[94] "Reddit Moves to Control Hate Speech and Misinformation in Two Forums."

[95] "Recommendation Algorithms and Extremist Content: A Review of Empirical Evidence."

[96] Rogers, "Why Reddit Banned Some Racist Subreddits But Kept Others."

[97] "Reddit Allows Hate Speech to Flourish in Its Global Forums, Moderators Say."

[98] "Reddit Moves to Control Hate Speech and Misinformation in Two Forums."

voices when it comes to breaking news and that it "rewards" trusted creators.[99] This an example of a positive intervention. Rather than just trying to target the problematic material, companies can instead try to elevate authoritative and trusted voices – for instance, many technology companies direct users towards credible resources like the Center for Disease Control website when users search for content related to the Covid-19 pandemic.[100]

While YouTube has ostensibly improved its recommendation algorithm to limit amplification or recommendation of TVEC, a 2022 study by Mozilla using 20,000+ participants found that YouTube's user controls (such as assigning "don't recommend this channel," "dislike," "remove from watch history") had very little impact on the videos recommended to users thereafter, which kept on resembling the initial video watched. However, the study was limited – as most literature on the subject tends to be – by the fact that there is no visibility into the inputs and design of the recommender system.[101] It is also unclear whether YouTube's experiment in improving its recommendation algorithm has been expanded to non-English-language media.[102]

Facebook (Meta) deprioritizes "misleading" content and uses positive interventions by providing users with information when viewing such content, though it is unclear the extent to which these policies are applied to TVEC. In countries like Myanmar, where Facebook has been criticized for failing to address misinformation inciting genocide, Meta states that it limits the spread of content shared by users with a history of sharing content in violation of Community Standards, a kind of de-prioritization, though it is unclear if this is an algorithmic process or not. Facebook also puts the onus on the user: it relies on users to be the end-moderators and provides options to flag or report TVEC.[103] It will reportedly remove or deprioritize "militarized social movements" (via lowering ranking in news feed and search and limiting recommendations of pages associated with said movements) as part of its Dangerous Individuals and Organizations policy.[104] Twitter engages in a similar approach, deprioritizing tweets with "downranks" to users (that are not direct followers) and will not recommend such tweets in its

[99] The YouTube Team, "Continuing Our Work to Improve Recommendations on YouTube."

[100] Geeng et al., "Social Media COVID-19 Misinformation Interventions Viewed Positively, But Have Limited Impact."

[101] "Hated That Video?"; "YouTube Regrets."

[102] Hern, "YouTube to Adjust UK Algorithm to Cut False and Extremist Content."

[103] "Understanding Social Media and Conflict."

[104] "An Update to How We Address Movements and Organizations Tied to Violence"; "Recommendation Algorithms and Extremist Content: A Review of Empirical Evidence."

"Top Search" function.[105] Other companies put the onus on the user to curate their feed more purposefully: for instance, Gab has three timeline options: *popular*, *controversial*, and *latest*.[106]

It is worth noting that having a policy and following through are two very different things. Right now, we lack the empirical data to know the extent to which adjustments to recommender systems are doing the job that they are meant to do (beyond observational and outcome-based data, interviews, and other often non-conclusive methods) or whether companies are following through on their promises to take down "violative content" or impose positive interventions. We also do not know the mechanism via which certain users or types of content are being deprioritized or de-platformed and whether this is a manual response or an algorithmic shift. The little we do know is self-reported by technology companies.

Overall, algorithmic adjustment can have immediate and broad impacts but tends to be limited to a single platform or application. For applications and platforms that have widespread usership and for which there are (arguably) few good alternative products, like Google's search function, the reach of algorithmic adjustment is substantial. For platforms where there are other alternatives, like social media platforms and forums, and messaging apps such as Facebook and WhatsApp, an unintended consequence of restricting content or deprioritization is pushing users to other platforms and entrenching echo chambers. This is not an argument against algorithmic deprioritization of TVEC for such companies; it is the opposite. If recommendation algorithms do indeed promote more extreme content, algorithmic adjustment away from this trend on major platforms would likely play an outsized role in stopping individuals who wouldn't otherwise seek out such information from being exposed to it and falling down the proverbial rabbit hole.

Algorithmic adjustment faces other issues in scalability. "Positive interventions" can be scaled quite easily and with few downsides. However, deprioritization and deplatforming are a can of worms when it comes to scaling up. First, we do not have a good idea as to how algorithms will work for every user, and thus adjustments must be carefully tailored to ensure a sufficiently low false positive rate. Second, it can be quite easy to evade algorithms using coded language and

[105] "Recommendation Algorithms and Extremist Content: A Review of Empirical Evidence."

[106] Whittaker et al., "Recommender Systems and the Amplification of Extremist Content."

other media manipulation techniques. Third, it runs into all sorts of freedom of speech contests and other ethical considerations. For instance, when applied by an authoritarian country with different standards of freedom of speech, the risk of censorship increases dramatically.

Recommendation algorithms sit on a continuum wherein any movement away from radicalizing effects is useful in combatting TVEC. Because they are often so influential, even small changes can have great impact. However, the impact on corporate bottom lines is a large disincentive for tech companies, and a rethinking of how to excise and reformulate current legal frameworks of obligations to maximize profit for shareholders would be helpful – a challenge for policymakers, rather than a technical issue.

5.4 **URL Sharing**

URL sharing is usually a manual process whereby a human conducting open-source intelligence (though sometimes web scrapers are used on known terrorist sites) will flag a URL that contains TVEC and will upload it to the Terrorist Content Analytics Platform (TCAP), an initiative started by Tech Against Terrorism (TaT) in late 2020. TaT sends URLs flagged as TVEC to technology companies. GIFCT has also partnered with TCAP to hash URLs and include some of them in its hash sharing database.[107]

In its first year of existence (December 2020-2021), TCAP sent 11,074 URLs to 65 registered tech companies, with 94% of flagged URLs removed.[108] This only scratches the surface, representing a very small number of total TVEC online. Monitoring and flagging of URLs also cannot keep up with submission rate of new TVEC posts given its manual nature.

In general, URL sharing is a rudimentary tactic that doesn't currently have a broad impact in terms of countering TVEC. It is as useful as manually flagging TVEC can be. However, TCAP's expanded cooperation with the GIFCT in URL sharing will increase the reach of flagged web pages and the use of this tool, and it may be

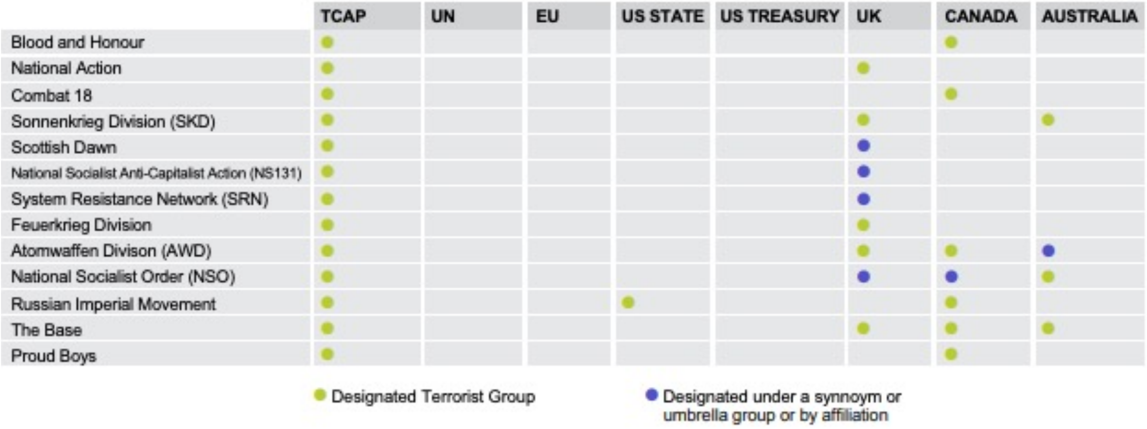
[107] "State of Play 2022: Trends in Terrorist and Violent Extremist Use of the Internet."

[108] "Transparency Report: Terrorist Content Analytics Platform."

helpful for smaller platforms that lack capacity for more sophisticated tools.[109]

More broadly, TaT's URL sharing efforts could play a role in norm setting, especially in creating a more inclusive definition of TVEC. TCAP's definition appears to be much broader than any other organization. *Figure 2* demonstrates that TCAP considers many more white supremacist groups to be terrorist groups than other entities. It also gives a sense of how narrowly focused the UN List is, and how wide variations exist among countries. Its inclusion of terrorist groups beyond Islamic extremism, coupled with its integration with powerful groups like GIFCT, could shift norms in favor of an expanded TVEC taxonomy.[110]

Figure 2. Infographic showing the far-right terrorist groups in scope of the TCAP and where they are currently designated.



	TCAP	UN	EU	US STATE	US TREASURY	UK	CANADA	AUSTRALIA
Blood and Honour	🟢						🟢	
National Action	🟢					🟢		
Combat 18	🟢						🟢	
Sonnenkrieg Division (SKD)	🟢					🟢		🟢
Scottish Dawn	🟢					🔵		
National Socialist Anti-Capitalist Action (NS131)	🟢					🔵		
System Resistance Network (SRN)	🟢					🔵		
Feuerkrieg Division	🟢					🟢		
Atomwaffen Divison (AWD)	🟢					🟢	🟢	🔵
National Socialist Order (NSO)	🟢					🔵	🔵	🟢
Russian Imperial Movement	🟢			🟢			🟢	
The Base	🟢					🟢	🟢	🟢
Proud Boys	🟢						🟢	

🟢 Designated Terrorist Group

🔵 Designated under a synnoym or umbrella group or by affiliation

[109] Tech Against Terrorism, "Gap Analysis and Recommendations for Deploying Technical Solutions to Tackle the Terrorist Use of the Internet."

[110] "Transparency Report: Terrorist Content Analytics Platform," 2021.

6. Discussion

This paper has identified some major barriers to progress in managing TVEC, despite only scratching the surface of the information ecosystem. Some key findings and recommendations are summarized below. We hope that they provide some guidance as to where policymakers, technology companies, and researchers could focus their efforts in the future.

1. A uniform and broader definition of TVEC should be formulated by policymakers and implemented across technology companies, to encompass specified *actions* or *activities* and *unaffiliated actors*, beyond just designated Islamic terrorist entities.

Of all the issues with current approaches to managing TVEC addressed in the paper, the most obvious first step towards future progress is to create a common and more inclusive definition of TVEC. Current definitions of TVEC are extraordinarily narrow, with dangerous implications. Most definitions tend to include only known Islamic terrorist groups, omitting a broad range of users and types of content. Not only does this mean that probably millions of TVEC data points are omitted from tech company efforts to identify and eliminate TVEC, but that we don't even *know* the extent of the problem because this data is not collected.

The definition question is a policy question, not a technical one, but it comes with a myriad of technical implications. Absent government policy on a definition of TVEC, the responsibility is falling on technology companies to draw the line, which often take the most conservative approach by adopting the narrow UN List as the benchmark. Technology companies themselves could drive normative change in defining TVEC as something broader than simply ISIS content, but so far progress has been slow, and is less likely to produce a uniform definition. No definition will be perfect, and it may be hard to differentiate between TVEC and non-TVEC when it comes to borderline content. But an imperfect yet improved definition is better than the status quo.

2. Not all technical tools are created equal. Multilateral and cross-company initiatives to combat TVEC should be inclusive of smaller firms and non-U.S. and non-European firms.

Hash-sharing is only as useful as it is scalable, within and between firms. ML tools require high upfront costs and ongoing resourcing. All these factors serve to make it harder for non-large Western tech companies to effectively apply the technical tools addressed in the paper.

3. Development of TVEC identification and management tools that are well-trained across different languages and cultural contexts is needed to ensure equitable standards in managing TVEC.

Current literature and multilateral/cross-company efforts to manage TVEC are heavily focused on Western companies and TVEC in English or Arabic, whereas a number of the world's largest technology companies are Chinese, and TVEC can be in many thousands of languages. Current tools to identify non-ISIS TVEC are learning and being fine-tuned using mostly English-speaking and data in a Western context, leaving behind other important contexts. Choosing the next targets of opportunity in machine learning translation will be a subjective exercise. Low hanging fruit include widely used online languages where ample training data are available for training ML translation tools, such as Chinese, Spanish, or Portuguese. Another approach could be to focus on languages that have been identified as adept at evading current identification tools, or on languages that are used in 'hot-spots' where online content in that language has been linked to real-life terrorism, such as Sinhala, German, or Norwegian.[111]

4. ML tools are not yet good enough to "algorithm our way out of the problem." A combination of tools must be employed. Establishing a standard of 'success' for ML tools vis-a-via TVEC could help to guide progress.

What does it mean for an ML detection tool to be successful? What is an

[111] Cai and Landon, "Attacks by White Extremists Are Growing. So Are Their Connections."; 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."

appropriate false positive rate? We don't yet have standards for these important questions. ML tools can pre-emptively address new TVEC but are not yet perfect and must be combined with other tools, which are also imperfect. Hash-sharing, while helpful, deals only with known TVEC, not new content. Meanwhile, human moderation and even human-automation synthesis is too slow to manage the constant stream of uploaded content. It can also be unreliable and inconsistent.

5. New legal measures to ingrain corporate social responsibility for technology companies may ease the burden of corporate obligation to shareholders – no more maximizing engagement and profit at any societal cost.

Even if technology companies want to manage TVEC, their incentive structure is legally geared towards prioritizing profit maximization for shareholders. Technology companies themselves may find it easier to deal with TVEC if assigned a legal responsibility to take care of this issue, rather than try to manage competing expectations from shareholders and governments/users.

6. Policymakers should be wary of unintended consequences of well-intentioned policies. Bad actors will always try to skirt the rules.

TVEC management is a chess game, not simple arithmetic. It can best be seen as a dynamic and iterative game whereby the opposition is sentient, varied, and constantly trying to find ways around new policies. New policies and tools to deal with TVEC challenges often come with unintended consequences that technologists and policymakers alike should bear in mind. For instance, de-platforming efforts have contributed to a proliferation of TVEC content on private message boards and in areas of the web that are harder to regulate. Users have turned to gaming platforms, alt-tech platforms, file hosting platforms, and video sharing platforms with more lax content policies, away from mainstream media.[112]

7. Public-private cooperation is critical in managing the threat of TVEC from a national security perspective.

As we learned after details of Operation GLOWING SYMPHONY were released

[112] "Gap Analysis and Recommendations for Deploying Technical Solutions to Tackle the Terrorist Use of the Internet."

to the public, national security agencies can be highly effective at eliminating terrorist threats online. Technology companies, as the hosts of TVEC, are also the gatekeepers of information relating to these actors. From a capabilities perspective, governments and technology companies have complementary know-how that, when combined, could result in more effective efforts to combat TVEC.

7. Conclusion

This paper has raised many more questions than it has answers. Clarifying the questions and understanding the issue is important before jumping to policy conclusions. While it has become a familiar trope, it remains critical that policymakers understand technical limitations, including awareness that even if technology companies want to solve the problem of TVEC, the assumed technical silver bullet may not exist. In particular, policy discussions on algorithmic transparency and ML tools need to be underlaid with sound technical understanding of recommendation and identification systems. Policy responses are currently oversimplified.

Similarly, a single tool or policy will not be enough to fix the problem. The complex nature of TVEC requires understanding and addressing all aspects of the problem on an individual basis and based on empirical evidence, rather than assumptions. This will require technology companies to be more transparent and will require governments to be more willing to learn about the technical limitations that companies face, to cooperate with technology companies, and take their interests into account. Innovation and policy making do not have to be at odds with one another. If approached prudently and cooperatively, regulation can support innovation, while also being in the public interest.

Finally, this paper should not be seen as an attempt to cover every issue relating to technical tools to combat TVEC; rather, it is an attempt to point to some of the main tools currently employed and identify some of the gaps in our capability and understanding. Encouragingly, groups like the GIFCT are proactive in seeking out research in areas where there is room for improvement, including some mentioned in this paper. Follow-up papers could investigate the following: new promising technical tools in development or possible changes to existing tools that lend promise for the future; how individual companies apply technical tools, from social media companies to video games to file sharing platforms; and whether companies at different levels of the technology stack could play different roles in implementing technical counter-TVEC tools.

Bibliography

"2022 GIFCT Transparency Report." Global Internet Forum to Counter Terrorism, December 2022.

"2022 GIFCT Transparency Report." Accessed February 20, 2023. https://gifct.org/wp-content/uploads/2022/12/GIFCT-Transparency-Report-2022.pdf.

Abbas, Ali, Max Senges, and Ronald A. Howard. "A Hippocratic Oath for Technologists." SSRN Scholarly Paper. Rochester, NY, October 1, 2018. https://papers.ssrn.com/abstract=3274327.

"An Overview of Perceptual Hashing | Journal of Online Trust and Safety." Accessed February 20, 2023. https://tsjournal.org/index.php/jots/article/view/24.

AP NEWS. "5 Months on, Christchurch Attacker Influences Others," April 20, 2021. https://apnews.com/article/australia-race-and-ethnicity-el-paso-new-zealand-mosque-attacks-tx-state-wire-e256dbf73bf043ec9ae49af18c4a33c3.

"Artificial Intelligence and Countering Violent Extremism: A Primer." Accessed February 20, 2023. https://gnet-research.org/wp-content/uploads/2020/10/GNET-Report-Artificial-Intelligence-and-Countering-Violent-Extremism-A-Primer_V2.pdf.

Avril Wong, Shubham Jain. "Deep Perceptual Hashing Is Not Robust to Adversarial Detection Avoidance Attacks." Text. https://cpg.doc.ic.ac.uk, August 11, 2022. https://cpg.doc.ic.ac.uk/blog/deephash-not-robust-to-detection-avoidance/.

Berger, J. M. "The ISIS Twitter Census: Making Sense of ISIS's Use of Twitter." Brookings (blog), November 30, 1AD. https://www.brookings.edu/blog/order-from-chaos/2015/03/06/the-isis-twitter-census-making-sense-of-isiss-use-of-twitter/.

Birnbaum, Emily. "TikTok Seeks to Join Tech Fight against Online Terrorism." Text. The Hill (blog), November 4, 2019. https://thehill.com/policy/technology/468884-tiktok-seeks-to-join-tech-fight-against-online-terrorism/.

Cai, Weiyi, and Simone Landon. "Attacks by White Extremists Are Growing. So Are Their Connections." The New York Times, April 3, 2019, sec. World. https://www.nytimes.com/interactive/2019/04/03/world/white-extremist-terrorism-christchurch.html, https://www.nytimes.com/interactive/2019/04/03/world/white-extremist-terrorism-christchurch.html.

Call, Christchurch. "Christchurch Call Text." Christchurch Call. Accessed April 22, 2023. https://www.christchurchcall.com/about/christchurch-call-text/.

Carr, Madeline. "Public-Private Partnerships in National Cyber-Security Strategies." International Affairs 92, no. 1 (January 2016): 43–62. https://doi.org/10.1111/1468-2346.12504.

"Case Study: Using the GIFCT Hash-Sharing Database on Small Tech Platforms." Tech Against Terrorism, 2018. https://www.counterextremism.com/sites/default/files/TAT%20--%20JustPaste.it%20GIFCT%20hash-sharing%20Case%20study.pdf.

China Briefing News. “China Passes Sweeping Recommendation Algorithm Regulations,” January 6, 2022. https://www.china-briefing.com/news/china-passes-sweeping-recommendation-algorithm-regulations-effect-march-1-2022/.

COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS The European Agenda on Security (2015). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52015DC0185.

“Content-Sharing Algorithms, Processes, and Positive Interventions Working Group.” Accessed April 5, 2023. https://gifct.org/wp-content/uploads/2021/07/GIFCT-CAPI1-2021.pdf.

“Corporations Don’t Have to Maximize Profits.” Accessed April 22, 2023. https://www.nytimes.com/roomfordebate/2015/04/16/what-are-corporations-obligations-to-shareholders/corporations-dont-have-to-maximize-profits.

Cottee, Simon. “Why It’s So Hard to Stop ISIS Propaganda.” The Atlantic, March 2, 2015. https://www.theatlantic.com/international/archive/2015/03/why-its-so-hard-to-stop-isis-propaganda/386216/.

Covington, Paul, Jay Adams, and Emre Sargin. “Deep Neural Networks for YouTube Recommendations.” In Proceedings of the 10th ACM Conference on Recommender Systems, 191–98. Boston Massachusetts USA: ACM, 2016. https://doi.org/10.1145/2959100.2959190.

Directive (EU) 2017/541 of the European Parliament and of the Council of 15 March 2017 on combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA, 088 OJ L § (2017). http://data.europa.eu/eli/dir/2017/541/oj/eng.

Du, Matt Sheehan, Sharon. “What China’s Algorithm Registry Reveals about AI Governance.” Carnegie Endowment for International Peace. Accessed April 5, 2023. https://carnegieendowment.org/2022/12/09/what-china-s-algorithm-registry-reveals-about-ai-governance-pub-88606.

“EU Internet Forum: Unified Action Needed in the Fight against Child Abuse Online and Terrorist Content Online.” Accessed April 22, 2023. https://home-affairs.ec.europa.eu/news/eu-internet-forum-unified-action-needed-fight-against-child-abuse-online-and-terrorist-content-2022-12-08_en.

European Commission - European Commission. “EU Internet Forum: A Major Step Forward in Curbing Terrorist Content on the Internet.” Text. Accessed February 20, 2023. https://ec.europa.eu/commission/presscorner/detail/en/IP_16_4328.

“European Union Internet Forum (EUIF).” Accessed April 22, 2023. https://home-affairs.ec.europa.eu/networks/european-union-internet-forum-euif_en.

“Facebook to Train AI Systems Using Police Firearms Training Videos,” September 17, 2019. https://www.cbsnews.com/news/facebook-to-train-ai-systems-using-police-firearms-training-videos/.

Federal Bureau of Investigation. “FBI Partnering with the Private Sector to Counter the Cyber Threat.” Speech. Accessed April 22, 2023. https://www.fbi.gov/news/speeches/fbi-partnering-with-private-sector-to-counter-the-cyber-threat-032222.

Fernandez, Alberto M. “Here to Stay and Growing: Combating ISIS Propaganda Networks.” U.S.-Islamic World Forum Papers 2015. The Brookings Project on U.S. Relations with the Islamic World. Center for Middle East Policy at Brookings, October 2015.

Finnemore, Martha, and Kathryn Sikkink. “International Norm Dynamics and Political Change.” International Organization 52, no. 4 (ed 1998): 887–917. https://doi.org/10.1162/002081898550789.

“Gap Analysis and Recommendations for Deploying Technical Solutions to Tackle the Terrorist Use of the Internet.” Accessed March 26, 2023. https://gifct.org/wp-content/uploads/2021/07/GIFCT-TAWG-2021.pdf.

Geeng, Christine, Tiona Francisco, Jevin West, and Franziska Roesner. “Social Media COVID-19 Misinformation Interventions Viewed Positively, But Have Limited Impact.” arXiv, December 20, 2020. http://arxiv.org/abs/2012.11055.

GIFCT. “Advances in Hashing for Counterterrorism.” GIFCT (blog), March 29, 2023. https://gifct.org/2023/03/29/advances-in-hashing-for-counterterrorism/.

———. “Broadening the GIFCT Hash-Sharing Database Taxonomy: An Assessment and Recommended Next Steps.” Global Internet Forum to Counter Terrorism, July 2021. https://gifct.org/wp-content/uploads/2021/07/GIFCT-TaxonomyReport-2021.pdf.

GIFCT. “Membership.” Accessed March 5, 2023. https://gifct.org/membership/.

“GIFCT Working Groups Output 2022.” Accessed February 20, 2023. https://gifct.org/wp-content/uploads/2022/09/GIFCT-22WG-Combined-US-Sizing2.1-2.pdf.

Goldsmith, Jack, and Tim Wu. Who Controls the Internet?: Illusions of a Borderless World. Oxford University Press, 2006. https://scholarship.law.columbia.edu/books/175.

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. “Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance.” Big Data & Society 7, no. 1 (January 1, 2020): 2053951719897945. https://doi.org/10.1177/2053951719897945.

Günther, Mario, and Atoosa Kasirzadeh. “Algorithmic and Human Decision Making: For a Double Standard of Transparency.” AI & SOCIETY 37, no. 1 (March 1, 2022): 375–81. https://doi.org/10.1007/s00146-021-01200-5.

Halliday, Josh. “Twitter’s Tony Wang: ‘We Are the Free Speech Wing of the Free Speech Party.’” The Guardian, March 22, 2012, sec. Media. https://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech.

Hao, Karen. “China May Be Chasing Impossible Dream by Trying to Harness Internet Algorithms.” Wall Street Journal, August 30, 2022, sec. World. https://www.wsj.com/articles/china-blazes-hazy-new-trail-to-tame-internets-algorithms-11661866321.

Hern, Alex. “YouTube to Adjust UK Algorithm to Cut False and Extremist Content.” The Guardian, August 27, 2019, sec. Technology. https://www.theguardian.com/technology/2019/aug/27/youtube-to-adjust-uk-algorithm-to-cut-false-and-extremist-content.

"Informal Meeting of the Heads of State or Government Brussels, 12 February 2015 - Statement by the Members of the European Council." Accessed April 22, 2023. https://www.consilium.europa.eu/en/press/press-releases/2015/02/12/european-council-statement-fight-against-terrorism/.

"Innovative Public Private Partnerships," n.d.

"Introduction to Locality-Sensitive Hashing." Accessed April 23, 2023. https://tylerneylon.com/a/lsh1/.

Jones, Dustin. "What Is the 'great Replacement' and How Is It Tied to the Buffalo Shooting Suspect?" NPR, May 16, 2022, sec. Race. https://www.npr.org/2022/05/16/1099034094/what-is-the-great-replacement-theory.

Khatib, Hadi Al, and Dia Kayyali. "Opinion | YouTube Is Erasing History." The New York Times, October 23, 2019, sec. Opinion. https://www.nytimes.com/2019/10/23/opinion/syria-youtube-content-moderation.html.

Lawfare. "Challenges in Combating Terrorism and Extremism Online," July 11, 2021. https://www.lawfareblog.com/challenges-combating-terrorism-and-extremism-online.

Legal Information Institute, Cornell Law School. "Terrorism." Accessed April 22, 2023. https://www.law.cornell.edu/wex/terrorism.

Llansó, Emma J. "No Amount of 'AI' in Content Moderation Will Solve Filtering's Prior-Restraint Problem." Big Data & Society 7, no. 1 (January 1, 2020): 2053951720920686. https://doi.org/10.1177/2053951720920686.

Lohr, Steve. "How Top-Valued Microsoft Has Avoided the Big Tech Backlash." The New York Times, September 8, 2019, sec. Technology. https://www.nytimes.com/2019/09/08/technology/microsoft-brad-smith.html.

Meta. "An Update to How We Address Movements and Organizations Tied to Violence," August 19, 2020. https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/.

Meta. "Meta Launches New Content Moderation Tool as It Takes Chair of Counter-Terrorism NGO," December 13, 2022. https://about.fb.com/news/2022/12/meta-launches-new-content-moderation-tool/.

Meta. "Understanding Social Media and Conflict," June 20, 2019. https://about.fb.com/news/2019/06/social-media-and-conflict/.

Miller, Greg, and Scott Higham. "In a Propaganda War against ISIS, the U.S. Tried to Play by the Enemy's Rules." Washington Post, April 9, 2023. https://www.washingtonpost.com/world/national-security/in-a-propaganda-war-us-tried-to-play-by-the-enemys-rules/2015/05/08/6eb6b732-e52f-11e4-81ea-0649268f729e_story.html.

MIT Technology Review. "Facebook Says It Has Removed 1.5 Million Copies of the New Zealand Terror Attack Video." Accessed April 22, 2023. https://www.technologyreview.com/2019/03/18/136587/facebook-says-it-has-removed-15-million-copies-of-the-new-zealand-terror-attack/.

MIT Technology Review. “Hated That Video? YouTube’s Algorithm Might Push You Another Just like It.” Accessed April 10, 2023. https://www.technologyreview.com/2022/09/20/1059709/youtube-algorithm-recommendations/.

MIT Technology Review. “How OpenAI Is Trying to Make ChatGPT Safer and Less Biased.” Accessed April 22, 2023. https://www.technologyreview.com/2023/02/21/1068893/how-openai-is-trying-to-make-chatgpt-safer-and-less-biased/.

MIT Technology Review. “The US Now Hosts More Child Sexual Abuse Material Online than Any Other Country.” Accessed April 22, 2023. https://www.technologyreview.com/2022/04/26/1051282/the-us-now-hosts-more-child-sexual-abuse-material-online-than-any-other-country/.

Nadeem, Reem. “Most Americans Think Social Media Sites Censor Political Viewpoints.” Pew Research Center: Internet, Science & Tech (blog), August 19, 2020. https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/.

Naughton, John. “To Err Is Human – Is That Why We Fear Machines That Can Be Made to Err Less?” The Observer, December 14, 2019, sec. Opinion. https://www.theguardian.com/commentisfree/2019/dec/14/err-is-human-why-fear-machines-made-to-err-less-algorithmic-bias.

OECD. “Transparency Reporting on Terrorist and Violent Extremist Content Online: An Update on the Global Top 50 Content Sharing Services.” Paris: OECD, July 15, 2021. https://doi.org/10.1787/8af4ab29-en.

OHCHR. “Moderating Online Content: Fighting Harm or Silencing Dissent?” Accessed April 23, 2023. https://www.ohchr.org/en/stories/2021/07/moderating-online-content-fighting-harm-or-silencing-dissent.

“Overview of Perceptual Hashing Technology.” United Kingdom Office of Communications (Ofcom), November 22, 2022.

Radsch, Courtney C. “GIFCT: Possibly the Most Important Acronym You’ve Never Heard Of.” Just Security, September 30, 2020. https://www.justsecurity.org/72603/gifct-possibly-the-most-important-acronym-youve-never-heard-of/.

“Recommendation Algorithms and Extremist Content: A Review of Empirical Evidence.” Accessed April 10, 2023. https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-TR-Empirical-1.1.pdf.

Reuters. “‘Lost Memories’: War Crimes Evidence Threatened by AI Moderation.” June 19, 2020, sec. Big Story 10. https://www.reuters.com/article/us-global-socialmedia-rights-trfn-idUSKBN23Q2TO.

Rogers, Kaleigh. “Why Reddit Banned Some Racist Subreddits But Kept Others.” Vice (blog), August 6, 2015. https://www.vice.com/en/article/539vv8/why-reddit-banned-some-racist-subreddits-but-kept-others.

Rowa, Dr. Jazz Yvonne. “The Contextuality of Lone Wolf Algorithms: An Examination of (Non)Violent Extremism in the Cyber-Physical Space.” Introducing 2022 GIFCT Working Group Outputs. Global Internet Forum to Counter Terrorism, 2022. https://gifct.org/wp-content/uploads/2022/09/GIFCT-22WG-ContextualityIntros-1.1.pdf.

Royal Commission of Inquiry into the Attack on Christchurch Mosques on 15 March 2019. “Assessment of the Individual and the Terrorist Attack.” Accessed April 11, 2023. https://christchurchattack.royalcommission.nz/the-report/firearms-licensing/assessment-of-the-individual-and-the-terrorist-attack/.

Royal Commission of Inquiry into the Attack on Christchurch Mosques on 15 March 2019. “General Life in New Zealand.” Accessed April 11, 2023. https://christchurchattack.royalcommission.nz/the-report/firearms-licensing/general-life-in-new-zealand/.

Royal Commission of Inquiry into the Attack on Christchurch Mosques on 15 March 2019. “The Report.” Accessed April 22, 2023. https://christchurchattack.royalcommission.nz/the-report/.

Saltman, Dr Erin. “Introducing 2022 GIFCT Working Group Outputs,” 2022.

Schmidt, Charles W. “ENVIRONMENT: California Out in Front.” Environmental Health Perspectives 115, no. 3 (March 2007): A144–47.

Schmitt, Eric. “U.S. Intensifies Effort to Blunt ISIS’ Message.” The New York Times, February 17, 2015, sec. World. https://www.nytimes.com/2015/02/17/world/middleeast/us-intensifies-effort-to-blunt-isis-message.html.

“Section 230 — Nurturing Innovation or Fostering Unaccountability?” U.S. Department of Justice, June 2020. https://www.justice.gov/file/1286331/download.

Shahani, Aarti. “With ‘Napalm Girl,’ Facebook Humans (Not Algorithms) Struggle To Be Editor.” NPR, September 10, 2016, sec. All Tech Considered. https://www.npr.org/sections/alltechconsidered/2016/09/10/493454256/with-napalm-girl-facebook-humans-not-algorithms-struggle-to-be-editor.

Smith, Brad, and Carol Ann Browne. Tools and Weapons: The Promise and the Peril of the Digital Age. Hodder & Stoughton, 2021.

“Social Responsibility and Enlightened Shareholder Primacy: Views From the Courtroom and Boardroom | Insights | Skadden, Arps, Slate, Meagher & Flom LLP.” Accessed April 22, 2023. https://www.skadden.com/insights/publications/2019/02/social-responsibility/social-responsibility-and-enlightened-shareholder.

Sottek, T. C. “Reddit Bans Several of Its Most Racist Communities.” The Verge, August 5, 2015. https://www.theverge.com/2015/8/5/9103393/reddit-content-policy-official.

“State of Play 2022: Trends in Terrorist and Violent Extremist Use of the Internet,” n.d.

Stiglitz, Joseph E., and Scott J. Wallsten. “Public-Private Technology Partnerships: Promises and Pitfalls.” American Behavioral Scientist 43, no. 1 (September 1999): 52–73. https://doi.org/10.1177/00027649921955155.

Tabuchi, Hiroko. “U.S. Climate Change Policy: Made in California.” The New York Times, September 27, 2017, sec. Climate. https://www.nytimes.com/2017/09/27/climate/california-climate-change.html.

Tech Against Terrorism. “Gap Analysis and Recommendations for Deploying Technical Solutions to Tackle the Terrorist Use of the Internet.” GIFCT Technical Approaches Working Group, Global Internet Forum to Counter Terrorism, July 2021. https://gifct.org/wp-content/uploads/2021/07/GIFCT-TAWG-2021.pdf.

Temple-Raston, Dina. “How The U.S. Hacked ISIS.” NPR, September 26, 2019, sec. I’ll Be Seeing You. https://www.npr.org/2019/09/26/763545811/how-the-u-s-hacked-isis.

“The Online Regulation Series | The European Union - Tech Against Terrorism,” October 19, 2020. https://www.techagainstterrorism.org/2020/10/19/the-online-regulation-series-the-european-union/, https://www.techagainstterrorism.org/2020/10/19/the-online-regulation-series-the-european-union/.

The YouTube Team. “Continuing Our Work to Improve Recommendations on YouTube.” YouTube Official Blog, January 25, 2019. https://blog.youtube/news-and-events/continuing-our-work-to-improve/.

Time. “Reddit Allows Hate Speech to Flourish in Its Global Forums, Moderators Say,” January 11, 2022. https://time.com/6121915/reddit-international-hate-speech/.

Time. “Reddit Moves to Control Hate Speech and Misinformation in Two Forums,” March 24, 2022. https://time.com/6160519/reddit-international-hate-speech-ban/.

“Transparency Report: Terrorist Content Analytics Platform.” Tech Against Terrorism, 2021. https://www.techagainstterrorism.org/wp-content/uploads/2022/03/Tech-Against-Terrorism-TCAP-Report-March-2022_v6.pdf.

“Transparency Report: Terrorist Content Analytics Platform.” Accessed March 5, 2023. https://www.techagainstterrorism.org/wp-content/uploads/2022/03/Tech-Against-Terrorism-TCAP-Report-March-2022_v6.pdf.

Tsai, Yi-Hsuan, and Ming-Hsuan Yang. “Locality Preserving Hashing.” In 2014 IEEE International Conference on Image Processing (ICIP), 2988–92, 2014. https://doi.org/10.1109/ICIP.2014.7025604.

Tubefilter. “YouTube Now Gets Over 400 Hours Of Content Uploaded Every Minute,” July 26, 2015. https://www.tubefilter.com/2015/07/26/youtube-400-hours-content-every-minute/.

UK Government. “Firearms Officers Begin Filming Training for Counter Terrorism Initiative.” Counter Terrorism Policing, October 24, 2019. https://www.counterterrorism.police.uk/firearms-officers-film-training/.

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: United Nations Office of Counter-Terorrism, 2021.

Whittaker, Joe. “Recommendation Algorithms and Extremist Content: A Review of Empirical Evidence.” Global Internet Forum to Counter Terrorism, July 2022. https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-TR-Empirical-1.1.pdf.

Whittaker, Joe. “Recommendation Systems and Extremism: What Do We Know?” GNET (blog), August 17, 2022. https://gnet-research.org/2022/08/17/recommendation-systems-and-extremism-what-do-we-know/.

Whittaker, Joe, Seán Looney, Alastair Reed, and Fabio Votta. “Recommender Systems and the Amplification of Extremist Content.” Internet Policy Review 10, no. 2 (June 30, 2021). https://policyreview.info/articles/analysis/recommender-systems-and-amplification-extremist-content.

“YouTube Regrets.” Mozilla Foundation, July 2021. https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf.

Science, Technology, and Public Policy Program
Belfer Center for Science and International Affairs
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138
www.belfercenter.org/stpp

Printed in the United States of America