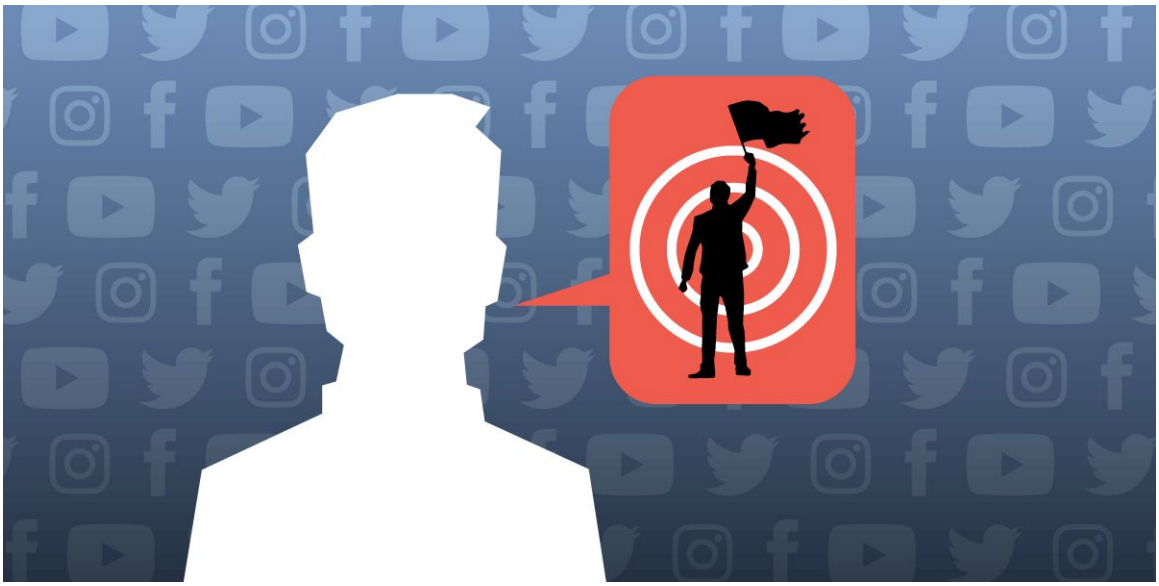

Caught in the Net:

THE IMPACT OF "EXTREMIST" SPEECH REGULATIONS ON HUMAN RIGHTS CONTENT

Abdul Rahman Al Jaloud, Hadi Al Khatib, Jeff Deutch, Dia Kayyali, and Jillian C. York

May 2019

ELECTRONIC FRONTIER FOUNDATION

SYRIAN ARCHIVE

WITNESS SEE IT FILM IT CHANGE IT

Contents

A joint publication of the Electronic Frontier Foundation, Syrian Archive, and Witness, 2019.
With assistance from: Hugh D'Andrade, Gennie Gebhart, David Greene, Jason Kelley

Introduction

Social media companies have long struggled with what to do about extremist content on their platforms. While most companies include provisions about "extremist" content in their community standards, until recently, such content was often vaguely defined, providing policymakers and content moderators a wide berth in determining what to remove, and what to allow. Unfortunately, companies have responded with overbroad and vague policies and practices that have led to mistakes at scale that are decimating human rights content.

The belief that deleting content on online platforms can solve the deeply rooted problems of extremism in modern society is a mistake. The examples highlighted in this document show that casting a wide net into the Internet with faulty automated moderation technology not only captures content deemed extremist, but also inadvertently captures useful content like human rights documentation, thus shrinking the democratic sphere. No proponent of automated content moderation has provided a satisfactory solution to this problem.

In recent years, following the rise of the Islamic State, companies have come under increasing pressure to undertake stricter measures when it comes to such speech. In the United States, this has come in the form of legislative proposals,[1] civil lawsuits from victims of terrorist attacks,[2] and pressure from the executive branch of the federal government.[3] The European Commission has ramped up its efforts from a code of conduct launched in 2017 that would require companies to review reported extremist content within 24 hours,[4] to a much more aggressive regulation that would create financial penalties for companies if they fail to act on extremist content within one hour.[5]

Regardless of these regulations, the vast majority of companies are already restrictive when it comes to extremist content. And because there is no globally agreed-upon definition of what constitutes a terrorist (and nations would inevitably disagree as to whether a specific entity met such a definition), U.S.-based companies such as Facebook, Twitter, and YouTube look to U.S. regulations to underpin their policies. As a

[1] Kelsey Harclerode, "Mandatory Reporting of User Content Chills Speech and Violates Privacy Rights," *Electronic Frontier Foundation*, 5 August 2015, https://www.eff.org/deeplinks/2015/08/mandatory-reporting-user-content-chills-speech-and-violates-privacy-rights.
[2] Aaron Mackey, "EFF to Court: Holding Twitter Responsible for Providing Material Support to Terrorists Would Violate Users' First Amendment Rights," *Electronic Frontier Foundation*, 8 June 2017, https://www.eff.org/deeplinks/2017/06/eff-court-holding-twitter-responsible-providing-material-support-terrorists-would.
[3] CBS News, "The Delicate Balance Fighting ISIS Online," 20 February 2015, https://www.cbsnews.com/news/world-governments-try-to-shut-down-isis-social-media-propaganda-operations/.
[4] Amar Toor, "Facebook, Twitter, Google, and Microsoft agree to EU hate speech rules," the *Verge*, 31 May 2016, https://www.theverge.com/2016/5/31/11817540/facebook-twitter-google-microsoft-hate-speech-europe.
[5] Colin Lecher, "Aggressive new terrorist content regulation passes EU vote," the *Verge*, 17 April 2019, https://www.theverge.com/2019/4/17/18412278/eu-terrorist-content-law-parliament-takedown.

result, the extremist groups that receive the most focus are typically those on the U.S. Department of State's list of Foreign Terrorist Organizations.[6] Facebook, for example, provides a list to moderators that includes photographs of leaders from groups on that list.[7]

But although companies use this list as guidance, they are not legally obligated under U.S. law to remove content that comes from these groups. As far as is publicly known, the U.S. government has not taken the position that allowing a designated foreign terrorist organization to use a free and freely available online platform is tantamount to "providing material support" for such an organization, as is prohibited under the patchwork of U.S. anti-terrorism laws. Although the laws prohibit the offering of "services" to terrorist organizations, the U.S. Supreme Court has limited that to concerted "acts done for the benefit of or at the command of another."[8] And U.S. courts have consistently rejected efforts to impose civil liability on online platforms when terrorist organizations use them for their communications.[9]

Content moderation and "extremist content"

Commercial content moderation is the process through which platforms make decisions about what content can and cannot be on their sites, based on their own Terms of Service, "community standards," or other rules. This process typically relies upon a system of community policing, whereby users of a given service report or "flag" content that they believe violates the rules. The content then enters a moderation queue, and a human moderator determines whether or not it violates the rules. Repeat violations on most platforms result in "punishments," in which a user is temporarily banned for increasing amounts of time.

Today, an increasing amount of moderation of extremist content is conducted through "automated flagging," a process in which platforms use their own proprietary tools to automatically detect potentially violating content that a human moderator then reviews. This often takes place before the content is ever seen by users.

In recent years, companies have exponentially increased their use of machine learning algorithms. In the computing world, algorithms are a set of instructions for doing something. Machine learning algorithms are algorithms that are given an initial set of data and some rules, and then learn and change as they come into contact with more data. In order to train a machine learning algorithm for this purpose, a company must create a dataset that includes a significant amount of content in one category, which they then feed to the algorithm for training. For example, in order to accurately identify extremist content, a company like YouTube would create a set of data that it defines as extremist—such as a large number of ISIS videos—then feed that data to its algorithm.

[6] U.S. Department of State, list of Foreign Terrorist Organizations, https://www.state.gov/foreign-terrorist-organizations/ (accessed 29 May 2019).
[7] The *Guardian*, "How Facebook Guides Moderators on Terrorist Content," 24 May 2017, https://www.theguardian.com/news/gallery/2017/may/24/how-facebook-guides-moderators-on-terrorist-content.
[8] Holder v. Humanitarian Law Project, 561 U.S. 1 (2010).
[9] See, for example, Fields v. Twitter, 881 F.3d 739 (9th Cir. 2018).

When an algorithm makes mistakes—as some of the cases below will illustrate—it can be difficult to understand why. Unless they are specifically designed to be "interpretable," machine learning algorithms cannot be understood by humans. Since they learn and grow on their own without the need for human input, we cannot see these algorithms' "thought processes." Unfortunately, platforms use machine learning algorithms that are proprietary and shielded from any external review. In fact, civil society and governments have been denied access to the training data or basic assumptions driving the algorithms, and there has never been any sort of third-party audit of such technology.

This problem has become more acute with the introduction of hashing databases for tracking and removing extremist content. Hashes are digital "fingerprints" of content that companies use to identify and remove content from their platforms. They are essentially unique, and allow for easy identification of specific content. When an image is identified as "terrorist content," it is tagged with a hash and entered into a database, allowing any future uploads of the same image to be easily identified.

Although such databases have been historically used to successfully track and remove child exploitation imagery from social media platforms, those databases—run by the National Center for Missing and Exploited Children (NCMEC) and the International Center for Missing and Exploited Children (ICMEC)—operate with oversight from law enforcement, and neither companies nor individual law enforcement officers have direct access to the images in the database.

The Global Internet Forum to Counter Terrorism (GIFCT), an industry initiative formed by Facebook, Microsoft, Twitter, and YouTube, used similar technology to introduce a database of "terrorist" images in 2016. The database is shared among GIFCT's member companies, which include smaller companies that do not have the resources to build their own databases. The result, however, is that a single database is used broadly across the Internet, and errors thus multiplied.

Much like the training content for extremist content algorithms, the GIFCT database is not shared with any members of civil society focused on human rights, and the GIFCT website offers minimal information about how it functions. And unlike the databases operated by NCMEC and ICMEC, the GIFCT database operates without external oversight. Instead, the determination of what constitutes extremism is left up to the companies.

To understand the scale of the issue, it is important to look at the numbers. Google's transparency report shows that YouTube removed 33 million videos in 2018,[10] amounting to roughly 90,000 per day. Of those flagged for potential violation of terms of service, 73% were removed through automated processes before the videos were available for viewing. Facebook removed roughly 15 million pieces of content deemed "terrorist propaganda" between October 2017 and October 2018. The company writes

[10] Google Transparency Report, "YouTube Community Guidelines enforcement," https://transparencyreport.google.com/youtube-policy/removals (accessed 12 May 2019).

that in the third quarter of 2018, Facebook found "99.5% of the content [was] subsequently removed before users reported it; the other 0.5% was reported by users first."[11] And Twitter, which famously took down 1.2 million terrorism-related accounts between 2015 and the final quarter of 2017,[12] removed an additional 166,153 accounts for terrorist content in the second half of 2018.[13]

Blunt measures affect marginalized users

Social media platforms wrongfully take down content across several different categories of speech, and no major social media company publishes its error rate. As a result, it can be difficult to understand where, when, and how often users suffer from inaccurate and mistaken takedowns. The examples below make clear, however, that it is difficult for human reviewers—and impossible for machines—to consistently differentiate activism, counter-speech, and satire about extremism from extremism itself. Blunt content moderation systems at scale inevitably make mistakes, and marginalized users are the ones who pay for those mistakes.

Example 1: Independence for Chechnya

In 2017, a Facebook group advocating for the independence of the Chechen Republic of Iskeria, called "Independence for Chechnya!", was mistakenly removed for violating the company's community standards barring "organizations engaged in terrorist activity or organized criminal activity," despite the fact that training manuals specifically identify the Chechen Republic of Iskeria as "not violating" the rules. A Facebook spokesperson said that the deletion was made in error and that the company "sometimes gets things wrong."[14]

Example 2: Kurdish Activism

Groups advocating for an independent Kurdistan are often the target of overbroad content moderation, despite the fact that only one such group—the Kurdistan Workers' Party (PKK)—is considered a terrorist organization by governments. As such, criticism and condemnation of the group is allowed on Facebook, but praise is not.[15]

---

[11] Facebook, "Community Standards Enforcement Report," https://transparency.facebook.com/community-standards-enforcement#terrorist-propaganda (accessed 12 May 2019).
[12] Don Reisinger, "Twitter Has Suspended 1.2 Million Terrorist Accounts Since 2015," *Fortune*, 5 April 2018, http://fortune.com/2018/04/05/twitter-terrorist-account-suspensions/.
[13] Foo Yun Chee, "Twitter suspended 166,153 accounts for terrorism content in second half 2018," *Reuters.* 9 May 2019, https://www.reuters.com/article/us-twitter-security/twitter-suspended-166153-accounts-for-terrorism-content-in-second-half-2018-idUSKCN1SF1LN.
[14] Julia Carrie Wong, "Facebook blocks Chechnya activist page in latest case of wrongful censorship," the *Guardian,* 6 June 2017, https://www.theguardian.com/technology/2017/jun/06/facebook-chechnya-political-activist-page-deleted.
[15] *The Guardian*, "How Facebook Guides Moderators on Terrorist Content." https://www.theguardian.com/news/gallery/2017/may/24/how-facebook-guides-moderators-on-terrorist-content

According to Human Rights Watch, Kurds have been frequent targets of human rights violations by the Turkish government.[16] The Turkish government is the world's worst jailer of journalists,[17] and is also a leader in censorship demands, requiring companies to take down anything illegal in the country, including criticism of the country's founder, Ataturk.

Kurdish activists have alleged that Facebook has repeatedly removed their posts that do not violate the platform's standards.[18] A Kurdish politician whose page was shut down called it a "dirty coalition" between the Turkish government and Facebook, noting that Turkey's ruling party, the AKP, has shared images of the leader of Hamas (which is also listed on the U.S. government's list of foreign terrorist organizations) with impunity.[19] When legitimate dissent such as that of Kurdish activists is removed by a company, either in error or as the result of government pressure, the company is effectively taking sides in a political dispute.

Example 3: Satirical Commentary

In 2017, Facebook removed an image posted by a prominent Emirati journalist of Hezbollah leader Hassan Nasrallah with a rainbow Pride flag overlaid on it.[20] The post, a commentary on Hezbollah's popularity amongst a certain segment of the political left despite a lack of support for LGBTQ rights, was too subtle for content moderators, who are directed to remove most images containing the faces of known terrorist leaders.[21]

Social media as evidence and memory

Social media documentation of human rights violations is critical for justice and accountability efforts, and in some cases it serves as collective memory. Videos and text posted online are living histories for some diaspora communities, and sometimes this documentation might offer the only evidence that a crime has been committed. Yet in too many cases, social media content moderation policies around extremism lead to the deletion of vital documentation. Restoring wrongfully deleted content is nearly

[16] Human Rights Watch, "Turkey, Events of 2017," https://www.hrw.org/world-report/2018/country-chapters/turkey (accessed 29 May 2019).
[17] Elana Beiser, "Hundreds of journalists jailed globally becomes the new normal," *Committee to Protect Journalists*, 13 December 2018, https://cpj.org/reports/2018/12/journalists-jailed-imprisoned-turkey-china-egypt-saudi-arabia.php.
[18] Sara Spary, "Facebook Is Embroiled In A Row With Activists Over "Censorship," *Buzzfeed*, 8 April 2016, https://www.buzzfeed.com/saraspary/facebook-in-dispute-with-pro-kurdish-activists-over-deleted.
[19] Hurriyet Daily News, Kurdish politicians to take action after Facebook admits to banning pages with PKK content," 29 April 2013, http://www.hurriyetdailynews.com/kurdish-politicians-to-take-action-after-facebook-admits-to-banning-pages-with-pkk-content-53465.
[20] Sophia Cope, Jillian C. York, and Jeremy Gillula, "Industry Efforts to Censor Pro-Terrorism Online Content Pose Risks to Free Speech," *Electronic Frontier Foundation*, 12 July 2017, https://www.eff.org/deeplinks/2017/07/industry-efforts-censor-pro-terrorism-online-content-pose-risks-free-speech.
[21] *The Guardian*, "How Facebook Guides Moderators on Terrorist Content." https://www.theguardian.com/news/gallery/2017/may/24/how-facebook-guides-moderators-on-terrorist-content

impossible if the person who posted the content is not alive, is arrested, or does not have access to email, all common issues in conflict zones.

YouTube censorship of conflict documentation in Syria, Yemen, and Ukraine

In Syria, human rights defenders predominantly use social media platforms to publish and publicise conflict documentation, and are able to use the medium effectively and often. In an interview with groups at the beginning of the Syrian uprising, activists said:

> We achieved a point when we realized we should start organizing ourselves, we should start something organized. Because all of the media channels refused to publish this kind of videos … There was actually no media coverage, only this one channel and social media, YouTube and Facebook ….Young people cooperated with channels, they made the channels actually, on YouTube …. These were the first local groups based on YouTube. They were organized, they had correspondents everywhere. They collected movies…. the first organized phenomenon in Syria was a media group.[22]

There are now more hours of social media content about the Syrian conflict than there have been hours in the conflict itself. With more than 50 videos still being uploaded each day, Syria presents arguably the first time in history that a conflict can be witnessed by anyone in the world, practically in real time.[23]

YouTube has used machine learning-powered automated flagging to terminate thousands of Syrian YouTube channels that were publishing videos documenting human rights violations. This includes channels from the Syrian Observatory for Human Rights, the Violation Documentation Center, Sham News Agency, and Aleppo Media Center. The terminated social media accounts ranged from documentation of protests in Syria to non-traditional media reporting on violent attacks, and did not incite violence or encourage dangerous activities[24]

At least 206,077 videos documenting rights violations were made unavailable on YouTube between 2011 and May 2019.[25] This includes 381 videos documenting airstrikes that targeted hospitals or medical facilities. One of those, titled "Tafas: Heavy artillery shelling of the national hospital 11/8/2012," showed Syrian government forces shelling medical staff and patients inside the hospital.[26]

---

[22] *Revolutionary Echoes from Syria*, (Hourriya, 2016), 18-21. Audio available at: https://archive.org/details/RevolutionaryEchoesFromSyria.
[23] Armin Rosen, "Erasing History: YouTube's Deletion Of Syria War Videos Concerns Human Rights Groups," *Fast Company*, 7 March 2018, https://www.fastcompany.com/40540411/erasing-history-youtubes-deletion-of-syria-war-videos-concerns-human-rights-groups.
[24] Syrian Archive, https://syrianarchive.org/en/tech-advocacy (accessed 29 May 2019).
[25] Ibid.
[26] "Tafas: Heavy artillery shelling of the national hospital 11/8/2012", YouTube video no longer available, https://www.youtube.com/watch?v=ipaaQGqtfTk.

Similar examples can be seen in Yemen, where since 2015, a war largely between the Saudi-led coalition and the Houthis has led to the direct killing of an estimated 70,000 people,[27] the displacement of over three million people,[28] and the death of an estimated 85,000 children from starvation.[29] A video titled "The first moment of bombing the big hall in Sana'a 08/10/2016"[30] and another video titled "Saudi massacre targeting displaced camps in Mazraq"[31] have both been made unavailable as a result of content moderation policies.

In Ukraine, which has been immersed in a war since the 2014 annexation of Crimea by Russia, videos on YouTube documenting the arming of pro-Russia and anti-government forces have also been removed. One example is titled "Military equipment supplied by Russia to the Donbass."[32]

Social media conflict evidence and case law

Content moderation hinders human rights efforts for legal accountability. Social media can offer irreplaceable evidence from conflict zones, particularly in places where foreign journalists, NGOs, and international monitoring agencies face difficulties accessing the country to document rights violations.[33] While courts and traditional documentation groups often lag behind in harnessing this potential, there is an emerging body of case law where social media content features prominently.

In 2016, for example, a Swedish court case was concluded against a former Syrian rebel who had taken part in the killing of seven captured Syrian soldiers. The court relied on content published on Facebook and Twitter to identify the time when and place where the soldiers were captured, as well as the fact that only 41 hours had passed between their capture and execution. Facebook was contacted by prosecutors in order to verify the content's metadata.[34]

In 2017, and again in 2018, the International Criminal Court (ICC) issued an arrest warrant for Libyan national Mahmoud Mustafa Busayf Al-Werfalli. Al-Werfalli was

[27] Al Jazeera, "More than 70,000 killed in Yemen's civil war: ACLED," 19 April 2019, https://www.aljazeera.com/news/2019/04/yemen-war-death-toll-reaches-70000-report-190419120508897.html.
[28] United Nations Agency for Refugees and Migrants, "More than 3 million displaced in Yemen – joint UN agency report," 22 August 2016, https://refugeesmigrants.un.org/more-3-million-displaced-yemen-%E2%80%93-joint-un-agency-report
[29] Sam Magdy, "Save the Children says 85,000 kids may have died of hunger in Yemen," *USA Today*, 21 November 2018, https://eu.usatoday.com/story/news/world/2018/11/21/yemen-children-hunger/2076683002/.
[30] "The first moment of bombing the big hall in Sana'a 08/10/2016", YouTube video no longer available, https://www.youtube.com/watch?v=dNax-YKBLNE.
[31] "Saudi massacre targeting displaced camps in Mazraq," YouTube video no longer available, https://www.youtube.com/watch?v=_Z5peAj4kTo.
[32] "Military equipment supplied by Russia to the Donbass," YouTube video no longer available, https://www.youtube.com/watch?v=CJm5bjM3Z5c.
[33] Elliot Higgins, "Weapons From the Former Yugoslavia Spread Through Syria's War," *The New York Times*, 25 February 2013, http://atwar.blogs.nytimes.com/2013/02/25/weapons-from-the-former-yugoslavia-spread-through-syrias-war/.
[34] Christina Anderson, "Syrian Rebel Gets Life Sentence for Mass Killing Caught on Video," *The New York Times*, 16 February 2017, https://www.nytimes.com/2017/02/16/world/europe/syrian-rebel-haisam-omar-sakhanh-sentenced.html.

accused of being directly responsible for the killing of 33 people, or for ordering the execution of those people. The arrest warrant states that evidence for seven of those incidents was based largely on video material and transcripts of video material posted to al-Werfalli's social media profiles.[35]

Conclusion

When tragedies like Christchurch happen, the impetus to respond is so great that it can lead to unintended outcomes. In fact, many of the same companies that have signed the Christchurch call have pushed back against legislative proposals for automated takedowns in front of policymakers in recent months. While these companies make big promises about automated content moderation in the news, they elsewhere admit that the technology is not foolproof. Facebook admits that even with human review, it has a high error rate.[36] Unfortunately, as the myriad examples above demonstrate, content moderation does not affect all groups evenly, and has the potential to further disenfranchise already marginalised communities.

The temptation to look to simple solutions to the complex problem of extremism online is strong, but governments and companies alike must not be hasty in rushing to solutions that compromise freedom of expression, the right to assembly, and the right to access information.

[35] The Prosecutor v. Mahmoud Mustafa Busayf Al-Werfalli, Case No.ICC-01/11-01/17-2, Warrant of Arrest (15 August 2017), https://www.icc-cpi.int/Pages/record.aspx?docNo=ICC-01/11-01/17-2.
[36] Ariana Tobin, Madeleine Varner, and Julia Angwin. "Facebook Uneven Enforcement of Hate Speech Rules Allows Vile Posts to Stay Up," *ProPublica*, 28 December 2017, https://www.propublica.org/article/facebook-enforcement-hate-speech-rules-mistakes.

About the Authors

Electronic Frontier Foundation

The Electronic Frontier Foundation is the leading nonprofit defending digital privacy, free speech, and innovation. Founded in 1990, EFF champions user privacy, free expression, and innovation through impact litigation, policy analysis, grassroots activism, and technology development. We work to ensure that rights and freedoms are enhanced and protected as our use of technology grows.

Syrian Archive

Syrian Archive is an organisation dedicated to preserving, memorialising and adding value to at-risk documentation related to human rights violations committed by all sides involved in the conflict in Syria. Syrian Archive uses this content to support justice and accountability efforts, and engages with social media platforms to assist documentation groups to recover or reinstate human rights content inadvertently removed due to content moderation policies.

WITNESS

WITNESS is an international nonprofit organization that uses education and advocacy to ensure human rights defenders and accidental witnesses use video ethically, safely, and effectively in their fight for human rights. WITNESS provides trainings and materials on video advocacy, video as evidence, archiving and more. WITNESS' Tech + Advocacy program focuses on the need to ensure that tech companies tools and policies support, rather than harm, human rights defenders through policy advocacy and platform accountability efforts.