Content Regulation in the Digital Age
Submission to the United Nations Human Rights Council
Special Rapporteur for Freedom of Expression
June 2018

Submitted by:
WITNESS
Contact: Dia Kayyali, Program Manager, tech + advocacy
dia@witness.org

WITNESS is a civil society organization that focuses on the use of video and technology to document and act on human rights abuses. Our headquarters are in New York City, but we have staff and partners all over the world. We collaborate with human rights defenders to share knowledge about how to safely, ethically, and effectively produce and utilize their own video, as well as archive and organize video created by others. Driven by our understanding of what is happening on the ground, we use our relationship with the tech world to enable greater freedom of expression and make existing tools and policies work better for human rights defenders and all users.

Summary: Today, the decisions made by technology companies like Facebook and YouTube/Google are affecting users' freedom of expression as much–if not more than–legislation and judicial decisions on. Content regulation has severely affected the ability of human rights defenders to share critical information, with little to no accountability as to how it occurs. We submit that companies must make a commitment to adhere to international human rights standards, including freedom of expression, even when it affects their financial bottom line or requires them to affirmatively defend attacks on rights by States.

***

Introduction: Despite sustained efforts by many civil society groups, content regulation, and in particular content removal, by major social media platforms remains an incredibly opaque process. In response to public pressure, the "big three"– Facebook, Google, and Twitter–have provided "transparency reports" around government requests for access to user data, but have not done so in any systematic way for content removal.[1] These companies, and other platforms are making content regulation decisions every day that respond to shareholders and governments and leave the free expression of all users–but especially human rights defenders–behind. Content regulation by companies is severely under-resourced and carried out in an inconsistent and unfair way. It continues to lack transparency, due process, real regional and local competency, and sufficient resources.

For the human rights defenders that WITNESS works with, this is having negative, and sometimes deadly, consequences. Content regulation is severely curtailing freedom of expression in an incredibly volatile and precarious time where human rights could either win widespread gains or continue to fall by the wayside. In the face of police, military and extremist violence in places like Syria, Brazil, and

1 "Companies tell us almost nothing about when they remove content or restrict users' accounts for violating their rules." *2017 Corporate Accountability Index,* Ranking Digital Rights, March 2017 *available at* https://rankingdigitalrights.org/index2017/assets/static/download/RDRindex2017report.pdf

Myanmar, human rights defenders are struggling to tell their stories to the world, and to anyone who can help them find accountability.

In our years of focusing on how technology companies' policies affect human rights defenders, we have seen a range of content regulation problems. We have looked at how companies take down graphic content that is shared for the purpose of exposing human rights abuses, while remaining unwilling to thoughtfully manage content such as videos of transphobic violence that are shared as entertainment.[2] We have seen social media platforms used as weapons against activists in some of the most vulnerable communities, where, with impunity, corrupt law enforcement or other bad actors create fake profiles that should be removed under terms of service, and use these profiles to harass and endanger the lives of activists.[3] Recently, we have seen a rapidly approaching danger that human rights defenders who are risking their lives to capture human rights abuses on the ground will be dismissed as "fake news." And we have seen how user-reporting based content regulation has allowed governments and bad actors to drive human rights defenders and marginalized groups like LGBTQ people off of platforms with coordinated reporting attacks.[4]

Most recently, we have seen the effect of unreasonable government demands to remove all "extremist content" almost as soon as it is posted.[5] This overzealous and shortsighted effort is having disastrous results on ongoing work that supports accountability for human rights abuses in Syria and other conflict zones–a result that will ultimately hamper efforts by bodies such as the offices of German and Swedish prosecutors, the International Criminal Court, and the International, Impartial and Independent Mechanism to hold extremists accountable through legal and policy processes.

---

2 Karen Stevenson with Kylar Broadus, *Capturing Hate: Eyewitness videos provide new source of data on prevalence of transphobic violence*, WITNESS Media Lab, Oct 2016, *available at* https://library.witness.org/product/capturing-hate-report/.

3 *See, e.g.,* Dia Kayyali, *Spies Use Tinder, and it's as Creepy As You'd Think*, Vice, 15 Nov 2016, https://motherboard.vice.com/en_us/article/78kdga/spies-use-tinder-and-its-as-creepy-as-youd-think; Felipe Larozza and Mauricio Fidalgo, *Na Favela da Maré, um Celular Põe mais Medo que um Fuzil*, Vice Brasil, 15 Apr 2015, https://www.vice.com/pt_br/article/nzjqq8/na-favela-da-mare-um-celular-poe-mais-medo-que-um-fuzil (in this instance, an ersatz profile for a group of human rights defenders which named and denounced drug traffickers in the name of that put their lives in danger almost immediately- Facebook did not take the page down until several stories were published in the press); we have also seen and addressed reports of this behavior directly with allies on the ground.

4 Ellery Roberts Biddle, "*We Will Choke You": How Indian Women Face Fatal Threats on Facebook While Trolls Roam Free,* Global Voices, 6 Aug 2015, https://advox.globalvoices.org/2015/08/06/we-will-choke-you-how-indian-women-face-fatal-threats-on-facebook-while-trolls-roam-free/

5 Nikolaj Nielsen, *Commission: 120 minutes to remove illegal online content* , EUobserver, 9 Jan 2018, https://euobserver.com/justice/140482; Laura Blanco and Jens-Henrik Jeppesen, *European Policymakers Continue Problematic Crackdown on Undesirable Online Speech*, Center for Democracy and Technology, 18 Jan 2018, https://cdt.org/blog/european-policymakers-continue-problematic-crackdown-on-undesirable-online-speech/

The content regulation process

What is content regulation?

It's important to be clear about what is meant by "content regulation." When WITNESS talks about content regulation, we are essentially referring to commercial content moderation, which is the process whereby companies determine what content, such as posts, videos, pictures, and other media, should and should not be allowed to be on their platforms. Content regulation should be the process whereby an existing and publicly available set of rules, generally in the form of clear terms of service, is applied in a consistent manner to content posted by all users, including government users. Unfortunately, this is not the case.

Methods of content regulation vary from site to site, but on the biggest platforms, historically content did not end up getting reviewed by human content moderators until it was reported by a human being. Now, artificial intelligence, in the form of computer vision and machine learning, has become a major part of content regulation, particularly content deemed as extremist by platforms and governments.

User reporting

In some cases, such as the enforcement of the policy that governs what names users can use on Facebook, platforms have maintained for many years that enforcement was exclusively user driven.[6] User reporting means that Facebook, Twitter, and YouTube/Google do not employ staff whose job it is to scour the sites looking for content that violate terms of service. Instead, each site has various ways to report content and accounts. In the case of its names policy, Facebook maintained that users could only end up in enforcement through user reporting. What is important to understand about user reporting is that it allows bad actors to coordinate and mass report accounts that it wants to disable: "Facebook groups have been created specifically for the purpose of reporting accounts, and in Vietnam government supporters have organized reporting sprees against political activists. Drag queens continue to notice specific groups being reported en masse,for example drag queens in a specific city that all know each other, or burlesque performers who've all performed at a specific show."[7] Human rights defenders, lacking similar resources, cannot defend themselves or respond offensively. It's logical to surmise that this is one of the tactics being used to silence activists who are speaking out about the Rohingya genocide.[8]

Taking humans out of the process: Machine learning and hashing

As noted above, the use of machine learning for content regulation has exponentially increased. Twitter, Facebook, and YouTube/Google have all made headlines for their use of machine learning to

---

6 Dia Kayyali, *Global Coalition to Facebook: 'Authentic Names' Are Authentically Dangerous for Your Users*, EFF, 5 Oct 2015, https://www.eff.org/deeplinks/2015/10/global-coalition-facebook-authentic-names-are-authentically-dangerous-your-users

7 Dia Kayyali, *Facebook's Name Policy Strikes Again, This Time at Native Americans*, EFF, 13 Feb 2015 https://www.eff.org/deeplinks/2015/02/facebooks-name-policy-strikes-again-time-native-americans

8 "[Rohingya activist Nay San] Lwin believes the removal of the posts is part of a campaign by government or government backers to discredit Rohingya online by reporting their posts to social media companies." from: BBC Trending, *Why are posts by Rohingya activists getting deleted?,* BBC, 23 Sep 2017, http://www.bbc.com/news/blogs-trending-41364633

moderate content, in particular images and videos that are considered extremist content.[9] In a June 2017 announcement Google said that it would "apply our most advanced machine learning research...to help us more quickly identify and remove extremist and terrorism-related content."[10] These systems work by feeding in to a model classifications of "good" and "bad" content, and then creating "classifiers"-algorithms that use these classifications to make decisions about content. Currently, YouTube notes that " 98 percent of the videos we remove for violent extremism are flagged by our machine-learning algorithms," Facebook states, " 99% of the ISIS and Al Qaeda-related terror content we remove from Facebook is content we detect before anyone in our community has flagged it to us, and in some cases, before it goes live on the site... Once we are aware of a piece of terror content, we remove 83% of subsequently uploaded copies within one hour of upload."[11]

One clear exception to human review is visual content that has been put through a process called hashing, where images are "convert[ed] into numerical values which are matched against databases of hashes from known illegal [or simply unwanted] images."[12] This process was originally developed for use against child exploitation images, but has also been applied to "extremist content," and these videos and images appear to either be prevented from getting uploaded or removed automatically. We have been working with companies for nearly a decade now to explain that human rights images are not just the image, but the surrounding context. For example, a video that includes images of violence committed by an extremist group, when used as a counter to propaganda showing staged good deeds committed by that same group, cannot simply be deleted. Unfortunately, this nuance may be lost in the process of hashing. A recent update from Twitter's blog about the Global Internet Forum to Counter Terrorism indicates that a database shared by companies now contains "more than 40,000 hashes."[13] As we note below, this is particularly concerning given that WITNESS has worked to restore thousands of videos on one platform, YouTube.[14] If these hashes have been shared even though they have been determined to be applied incorrectly after the fact, the effect of such incorrect content regulation will spread like a disease from platform to platform.

State involvement

In some cases, platforms will restrict access to content in compliance with local law. This is problematic both because it is not always clear how they decide which laws to comply with, even from their "transparency reports." Furthermore, cooperation through bodies like the Global Internet Forum

---

9 Josh Constine, *Facebook spares humans by fighting offensive photos with AI,* TechCrunch, 31 May 2016, https://techcrunch.com/2016/05/31/terminating-abuse/

10 Kent Walker (General Counsel, Google) *Four steps we're taking today to fight terrorism online*, Google Blog, 18 June 2017 https://www.blog.google/topics/google-europe/four-steps-were-taking-today-fight-online-terror/

11 Monika Bickert and Brian Fishman, *Hard Questions: Are We Winning the War On Terrorism Online?* Facebook Newsroom, 28 Nov 2017 https://newsroom.fb.com/news/2017/11/hard-questions-are-we-winning-the-war-on-terrorism-online/

12 *See, Microsoft PhotoDna,* (accessed 20 Jan 2018), https://www.microsoft.com/en-us/photodna

13 Twitter Public Policy, *Update on the Global Internet Forum to Counter Terrorism* Twitter Blog, 4 Dec 2017, https://blog.twitter.com/official/en_us/topics/events/2017/GIFCTupdate.html

14 Dia Kayyali and Raja Althabani, *Vital Human Rights Evidence in Syria is Dissapearing from YouTube*, Aug 2017 WITNESS https://blog.witness.org/2017/08/vital-human-rights-evidence-syria-disappearing-youtube/

to Counter Terrorism, while at surface level a positive move, provides avenues for government definitions of extremist content to permeate content regulation.

While State designations of terrorism may be uncontroversial in some cases, it is not always so- for example, Foreign Policy revealed in October 2017 that the Federal Bureau of Investigation had created a new class of domestic terrorists: "black identity extremists."[15] A former senior Department of Homeland Security official told Foreign Policy, "This is a new umbrella designation that has no basis. There are civil rights and privacy issues all over this."And civil rights activists Malkia Cyril and Shanelle Matthews note the stark parallels between this designation and the now-reviled program of FBI repression against civil rights leaders in the 1960s known as COINTELPRO.[16] Given that Black Lives Matter is inextricably linked to hashtags and social media, it's disturbing to imagine the consequences if Facebook, Twitter, and YouTube/Google are taking this designation seriously.

All of this is also complicated by the fact that some platforms acknowledge direct relationships with repressive States. For example, in Cambodia, where freedom of expression has come under serious attack in recent years, the government brags about having a close relationship with Facebook.[17] Posts critical of the Prime Minister are sent directly to Facebook, rather than ever having to go through the reporting process most people are subjected to, and Facebook "often complies." In fact, the Prime Minister's social media team "does their best to exploit Facebook's own rules." This tactic has become familiar to human rights defenders around the world, who often rely on Facebook to spread their message and may break rules such as real name policies due to the very real threat of being jailed or murdered for daring to have an opinion.

Furthermore, there are myriad cases that indicate that in the right circumstances technology companies will push back on legal frameworks.[18] We applaud efforts by platforms to stand up for human rights such as privacy and freedom of expression, but instead of such a piecemeal approach, companies should clearly commit themselves to defending international human rights standards, even when doing

---

[15] Sharon Weinberger and Jana Winter, *The FBI's New U.S. Terrorist Threat: 'Black Identity Extremists',* Foreign Policy, 6 Oct 2017, http://foreignpolicy.com/2017/10/06/the-fbi-has-identified-a-new-domestic-terrorist-threat-and-its-black-identity-extremists/

[16] Malkia Cyril and Shanelle Matthews, *We say black lives matter. The FBI says that makes us a security threat.*, Washington Post, 19 Oct 2017, https://www.washingtonpost.com/news/posteverything/wp/2017/10/19/we-say-black-lives-matter-the-fbi-says-that-makes-us-a-security-threat

[17] Megha Rajagopalan, *This Country's Democracy Has Fallen Apart — And It Played Out To Millions On Facebook*, Buzzfeed, 21 Jan 2018, https://www.buzzfeed.com/meghara/facebook-cambodia-democracy

[18] *See, e.g.,* Reuters *French court refers 'right to be forgotten' dispute to top EU court* Reuters, 19 Jul 2017, https://www.reuters.com/article/us-google-litigation/french-court-refers-right-to-be-forgotten-dispute-to-top-eu-court; James Titcomb, *Twitter drops lawsuit against US government as order to unmask anti-Trump account withdrawn,* Telegraph UK, 7 Apr 2017, http://www.telegraph.co.uk/technology/2017/04/07/twitter-sues-us-government-order-unmask-anti-trump-account/

so might affect their financial bottom line or make it difficult to operate offices within the borders of countries hostile to freedom of expression.

What's wrong with content regulation now?
There are myriad problems with content regulation now, including lack of transparency around the processes used at each company; overzealous and opaque use of machine learning; under-resourcing of content moderation and user outreach; and a continued lack of cultural, linguistic, and political fairness and competency.

The ramped-up use of machine learning in particular has caused problems for our allies on the ground. We have seen content removed that was created by extremists, but is being used to demonstrate and catalogue human rights abuses. We have also seen a continued misunderstanding from platforms that videos *about* violence are not necessarily *promoting* violence. We have also seen myriad instances of content with no connection to extremism being improperly removed under that rubric. In particular, we have seen hundreds of videos with no conceivable connection to extremism that do contain graphic content that have been improperly removed. According to YouTube's own help center, such content should be allowed on YouTube for "educational, documentary, scientific, or artistic purpose[s]".[19] Surely, documenting human rights abuses falls under that exception–and historically, at YouTube, it has.

Removal of videos from Syria
In July 2017–less than one month after Google announced that it would be using machine learning–our partner, the Syrian Archive, started hearing about removals of videos that were incredibly valuable as human rights evidence.[20] Some of these were videos created by groups such as ISIL, that were being collected by "aggregators" like the Violations Documentation Center (VDC). The VDC was established in 2011 by some of Syria's most reputable human rights advocates, with the aim of "ensur[ing] careful and independent documentation all violations and crimes in Syria."[21] We met VDC for the first time in 2012 following the exponential increase in video content coming out of Syria. Like many other groups working in Syria, they sought support around ensuring the effective, safe and ethical use of video documentation coming out of the country. VDC's content, like that of many others whose channels have been taken down, has since supported the reporting and advocacy efforts of major media outlets, human rights organizations and independent investigative bodies, including reports to the United Nations Human Rights Council.[22]

We quickly joined with the Archive to try to remedy these deletions. Videos that clearly documented human rights abuses in Syria, as well as the channels that featured these videos and existed for the

---

19 YouTube, *Policies, safety, and reporting > help center, (accessed 21 Dec 2017)* https://support.google.com/youtube/answer/2802008

20 *See, Survey of social media channels removed by new @YouTube machine learning and restored w/ help by @witnessorg,*" posted by Syrian Archive on Twitter, 5 Sep 2017 *at* https://twitter.com/syrian_archive/status/905019172370960384

21 Violations Documentation Center, *Our Story*, (accessed 15 Dec 2017), http://vdc-sy.net/our-story/

22 UN Human Rights Council, *Report of the independent international commission of inquiry on the Syrian Arab Republic*, 22 Feb 2012, A/HRC/19/69, *available at* http://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session19/A-HRC-19-69_en.pdf

purpose of telling this important story to the world, were being removed by YouTube at an astonishing pace—at the time, the Archive put the number between 120-150 thousand, and we continue to document removals.[23] This is true even as YouTube has worked closely with us to understand removals and restore channels. We are seeing some prominent channels with a well-documented history of providing quality videos, such as the Shaam Network, that have had their channel terminated up to five times. Shaam Network is one of the most important sources of videos documenting the conflict, with tens of thousands of videos at stake each time their channel is removed. We suspect that this is a function of video hashing, where a video has been improperly removed and then triggers repeated removals due to failure to correct in the system. Under YouTube's "three strikes" system, this could lead to multiple suspensions. These removals are hampering some of the best documentation of war crimes and human rights abuses in Syria.

What's more, we have seen myriad videos removed that could not conceivably have a link to extremist content. For example, the Syrian Archive showed us a video that has been removed multiple times of an attack on civilians in Syria. The news-style video shows rubble and a body being zipped up into a body bag, and it is clearly reporting on the situation as newsworthy. We have not received any explanation as to why the video has been removed. It appears to be a misunderstanding of YouTube's exception to its rule against posting graphic content.

The utility of these videos is not hypothetical. WITNESS has pioneered resources for activists who are creating video as evidence, culminating in the 2016 release of our Video as Evidence guide.[24] We've now seen prosecutors, courts, and other bodies that initially struggled with such evidence start to understand how to appropriately use it. In a historic first, on August 15 of this year, the International Criminal Court issued a warrant for the arrest of Mahmoud Mustafa Busayf Al-Werfalli for crimes in Libya, based partially on 7 videos obtained on social media.[25] A recent Human Rights Watch report explains that Swedish and German authorities are conducting structural and specific investigations into serious international crimes committed in Syria; authorities in Sweden are conducting 13 investigations against specific individuals for crimes in Syria and German authorities are conducting 27 investigations against specific individuals for grave crimes committed in Syria and Iraq.[26] As Human Rights Watch points out, "publicly available platforms, such as social media" and "various nongovernmental documentation groups working beyond their borders" such as the Violations Documentation Center and the Syrian Archive are essential in these prosecutions. These prosecutions are, unfortunately, focused on terrorism-related charges instead of war crimes prosecutions against Assad's regime-partly because

---

[23] Alex MacDonald, *YouTube AI deletes war crime videos as 'extremist material'*, Middle East Eye, 9 Aug 2017, http://www.middleeasteye.net/news/youtube-criticised-after-middle-east-video-taken-down-over-extremist-content-1244893230

[24] WITNESS, *WITNESS Launches Video as Evidence Field Guide for Citizens, Activists, Lawyers,* 30 Mar 2016, https://witness.org/witness-launches-video-as-evidence-guide-for-citizens-activists-lawyers/

[25] International Criminal Court Warrant of Arrest in *The Prosecutor v. Mahmoud Mustafa Busayf Al-Werfalli,* issued 15 Aug 2017 and *available at* https://www.icc-cpi.int/Pages/record.aspx?docNo=ICC-01/11-01/17-2

[26] Human Rights Watch, *"These are the Crimes we are Fleeing": Justice for Syria in Swedish and German Courts*, 3 Oct 2017, *available at* https://www.hrw.org/report/2017/10/03/these-are-crimes-we-are-fleeing/justice-syria-swedish-and-german-courts

of ISIS' predilection for filming itself committing human rights abuses as propaganda. However, they are the only real judicial movement towards justice for Syrians. The current trend of indiscriminately deleting "extremist content," regardless of the reason and manner in which this content is posted, imperils these investigations.

What needs to change

Invest in meaningful, regular outreach to human rights defenders and targeted outreach in key situations

Companies should not be waiting for the United Nations or civil society organizations to explain the human rights effects of their tools, products, and policies. Technology companies today need to learn a lesson from the movement for environmental protection. It was not until activists started speaking out about the incredible damage being done by industry that the United States and other countries truly began to regulate companies.[27] But by that point, potentially irreparable damage had been done. It appears we are almost at that point already when it comes to freedom of expression. But there is still time. Companies do not have to wait for the technological version of Love Canal to take the unintended effects of their activities seriously. Love Canal was a community built on top of a chemical waste site in the 1950s. In 1978, a record amount of rainfall exposed long-buried chemicals. An observer in 1978 wrote:" Everywhere the air had a faint, choking smell. Children returned from play with burns on their hands and faces.And then there were the birth defects. The New York State Health Department is continuing an investigation into a disturbingly high rate of miscarriages, along with five birth-defect cases detected thus far in the area."[28]

Instead of waiting to hear about how content regulation is rewarding the perpetrators of human rights abuses or silencing voices that may be relying on the public pressure created by social media to save their lives, companies can invest some of their huge profits in paying human rights defenders for their time and understanding how they use and experience technology. When it becomes clear that something is wrong—such as the silencing of Rohingya activists—platforms can invest in targeted outreach to affected communities and take immediate steps to fix the problem, such as hiring content moderators whose only job is to ensure that marginalized communities are not falling victim to targeted attacks. These are achievable standards, demonstrated by the fact that companies are already doing some of this. Some companies are doing a better job than others, but it isn't enough. The "big three" in particular need to invest more now.

Constantly assess machine learning and AI deployment and address errors

As social media experts Catherine Buni and Soraya Chemaly write, "Facebook, like Twitter, YouTube, and other similar platforms, refuse to share details of their content regulation. While there are many understandable reasons — abusers gaming the system, competitors gaining insights into internal processes — there are many more compelling ones arguing for more transparency and accountability.[29]

---

[27] *See,* Environmental Protection Agency, *EPA History* (as available on 11 Nov 2016) https://web.archive.org/web/20161111223737/https://www.epa.gov/history;

[28] Eckardt Beck, *The Love Canal Tragedy*, EPA Journal, Jan 1979 *available at* https://web.archive.org/web/20170424085641/https://archive.epa.gov/epa/aboutepa/love-canal-tragedy.html

[29] Catherine Buni and Soraya Chemaly, *How Do You Fix Facebook's Moderation Problem? Figure Out What Facebook Is*, The Verge, 25 May 2017, https://www.theverge.com/2017/5/25/15690590/facebook-leak-report-moderation-broken

This is even more applicable to machine learning processes. Human errors can be understood, but that is not always the case with machine learning. And once a model has been fed incorrect data, if not corrected for, that data will cause further mistakes. For example, as noted above, errors in the Global Internet Forum to Counter Terrorism hash database will continue to propagate if not corrected. And even if improperly added images are removed, it seems as though they will already have affected machine learning processes. While we don't necessarily agree that protecting commercial interests is a good reason to curtail freedom of expression, even within that framework companies can do better. In the same way that companies pay for security and financial audits, they can audit their machine learning processes. There is a growing field of scholarship in this area. In fact, the "Fairness, Accountability, and Transparency in Machine Learning Processes (FAT/ML)" organization includes, in addition to academics, representatives from Google, Microsoft, and Cloudflare.[30] Recently, FAT/ML published a set of "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms," that provide a good set of questions for assessing and auditing algorithms.[31] Though it is commendable that a Google employee helped to create these principles, the company itself does not appear to have applied these principles to YouTube's content moderation algorithms.

Hire more staff and provide them with sufficient time and resources

In recent months, big companies have made myriad announcements about hiring more content moderators.[32] This is a positive step. What is less clear, however, is whether all of these staff are given sufficient support and time to genuinely assess the content they are reviewing.[33] Instead of simply hiring more staff, companies need to allow content moderators sufficient time to genuinely review material, rather than requiring them to move through a large volume of content in a restricted environment.[34] Furthermore, they need to ensure that they are providing content moderators, whether they are direct staff or consultants, with sufficient mental health resources and training. They also need

---

30 FAT/ML, *Organization*, (accessed 20 Jan 2018), https://www.fatml.org/organization

31 FAT/ML, *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms*, (accessed 20 Jan 2018), https://www.fatml.org/resources/principles-for-accountable-algorithms

32 *See, e.g.*, Samuel Gibbs, *Facebook Live: Zuckerberg adds 3,000 moderators in wake of murders,*The Guardian, 3 May 2017, https://www.theguardian.com/technology/2017/may/03/facebook-live-zuckerberg-adds-3000-moderators-murders; Susan Wojcicki, CEO of YouTube, *Expanding our work against abuse of our platform,* YouTube Official Blog*,* 4 Dec 2017, https://youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of-our.html

33 *See, e.g., comments from content moderators:* "We were underpaid and undervalued," said the man, who earned roughly $15 per hour removing terrorist content from the social network after a two-week training course…Every day people would have to visit psychologists. Some couldn't sleep or they had nightmares." from: Olicia Solon, *Underpaid and overburdened: the life of a Facebook moderator*, The Guardian, 25 May 2017, https://www.theguardian.com/news/2017/may/25/facebook-moderator-underpaid-overburdened-extreme-content

34 "He has only a few seconds to decide. New posts are appearing constantly at the top of the screen, pushing the others down." from: Adrian Chen, *The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed,* Wired, 23 Oct 2014, https://www.wired.com/2014/10/content-moderation/

to ensure that content moderators are genuinely screened for political bias.[35] This will be expensive, but that needs to be considered the cost of doing business.

Include real information about content regulation in transparency reports
For years, civil society has been calling on companies to provide information about content regulation in their transparency reports. Currently, the "big three" only provide minimal information. We join with Ranking Digital Rights, Online Censorship, and myriad other voices in calling on companies to provide clear information about content regulation in their transparency reports. That means clearly stating what content is removed, how much is removed, what the justifications for removal are, what the relationship between State requests and removals is, and how many of these removals are being conducted exclusively by or with the aid of machine learning.

Provide due process and appeals for users
Both the process of content regulation and the process of appealing content regulations continues to demonstrate a severe lack of due process. We have seen dozens of messages sent to Facebook and YouTube users telling them that they have violated the terms of service that simply do not provide sufficient information about exactly how they violated the terms and why their media was taken down. Furthermore, when sites such as YouTube shut down entire channels based on their three strikes system, they do not provide information about what the other strikes were, which can cause confusion for users. Users need to be provided with more tailored messages and greater human interaction, preferably in the form of an email address to an individual employee, instead of canned messages that do not explain the problem- and as noted above, in critical situations companies should reach out to human rights defenders to ensure they understand why their content is being regulated. Users also need to be provided with secure channels for appeals, rather than being asked to provide sensitive information such as photos of id cards over open channels.

Push back on laws or requests that violate human rights
When confronted about some content removal, companies complain that they are "only complying with the law." This is problematic both because it is not always clear how they decide which laws to comply with, even from their "transparency reports" and because, as noted above, companies have gone to court in response to State demands, as well as joining policy coalitions in the United States and elsewhere. Instead of such a piecemeal approach, companies should clearly commit themselves to international human rights standards—even in places where it might affect their financial bottom line.

Conclusion
Proper content regulation is not a simple matter, but there is no question that with a commitment to human rights, it could be better. We hope this submission, and the ensuing report, will help companies understand how

[35] *See, eg, complaints by Rohingya activists about moderation by Facebook:* "'They should hire a team who are unbiased,' he said. 'This team is completely biased.'" from: Betsy Woodruff, *Exclusive: Facebook Silences Rohingya Reports of Ethnic Cleansing,* The Daily Beast, 18 Sep 2017, https://www.thedailybeast.com/exclusive-rohingya-activists-say-facebook-silences-them