Submission

Content Regulation in the Digital Age

2 February 2018

Table of Contents

1 Company compliance with State laws....1
2 Other State Requests....2
3 Global removals....2
4 Individuals at risk....3
5 Content regulation processes....3
6 Bias and non-discrimination....4
7 Appeals and remedies....5
8 Automation and content moderation....5
9 Transparency....6
10 Examples....6

The Electronic Frontier Foundation (EFF) is the leading nonprofit organization defending civil liberties in the digital world. 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. We thank you for the opportunity to provide input into your work on the role of Internet platforms with respect to content regulation in the digital age.

1 Company compliance with State laws

In considering whether an Internet company should comply with national content regulation laws, EFF's general starting point is that companies should only comply with such laws if they have

“boots on the ground” (that is, a physical presence) in the given country, since failing to do so could result in enforcement measures being taken against the company, including the possible prosecution of its staff who are working there.

But there are exceptions to this general working rule. Even if a company does not have physical presence in a country, it may choose to comply with that country’s content regulation laws if these are consistent both with its terms of service, and with its responsibility to respect human rights. For example, if a platform’s terms of service prohibit hateful speech, and if a country in which such speech is also unlawful requests its removal from that platform, it would be acceptable for the company to honor such a request, provided that it is transparent about doing so and follows due process.[1]

Conversely, even if a company does have a physical presence in a country, there are cases in which it should not comply with national laws, and should instead resist them by lawful means such challenging them in court, or even by withdrawing its operations from that country, if this is necessary in order to fulfill its responsibility to respect international human rights standards.[2]

In order not to become subject to national laws that would require it to violate human rights, a company should consider very carefully what jurisdictions it should enter and establish a physical presence in. If a company has a choice to establish a regional presence in one of a number of countries, it should seriously consider establishing its office in the country that has the best human rights record, from where it can offer service to neighboring countries without running the risk of becoming subject to their laws.

2 Other State Requests

Companies receiving requests from governments for content removals under their terms of service should enumerate these requests in their transparency reports separately from orders for the removal of illegal content, and separately from the requests that are received from non-State actors. For example, Twitter now separately identifies whether a government request for removal of content is based upon Twitter’s own terms of service, or is based on a violation of law. This is an emerging best practice in transparency reporting, which we encourage other Internet platforms to adopt.

3 Global removals

Where a national law does require a company to restrict content and where a company does so, where technically and legally possible it should use geolocation techniques to limit the restriction of content to the geographical borders of the country in question.[3]

When a national court order from a jurisdiction where a company has physical presence requires the global removal of content, the company has to weigh its obligations towards that jurisdiction, against its obligations to uphold the human rights of its users in other jurisdictions—and, perhaps,

1 Manila Principles on Intermediary Liability, paras 3(f) and 5.

2 One impetus for the formation of the Global Network Initiative (GNI) in 2008 was the 2002 arrest of Chinese democracy activists whose identities were disclosed to authorities by Yahoo!. The GNI, it was hoped, would facilitate the development of higher standards of protection for freedom of expression and privacy by American tech companies, and hold its members to those standards.

3 Manila Principles on Intermediary Liability, para 4(c).

its conflicting legal obligations from other jurisdictions.

A current case in point is Google's response to the Equustek litigation, wherein it was ordered by the Supreme Court of Canada to remove links to search content that allegedly infringed Canadian trade secrets law from its global search index. Google sought and received a preliminary injunction from a U.S. Federal Court preventing the enforcement of that order against Google in the United States, on the grounds that Google's activities in indexing such content were protected under U.S. law by Section 230 of the Communications Decency Act.[4]

4 Individuals at risk

Online harassment can be profoundly damaging to the free speech and privacy rights of the people targeted. It is frequently used to intimidate those with less political or social power, and affects some groups disproportionately, including women and racial and religious minorities. That means that not everyone appreciates the level to which it negatively affects the lives of others.

We oppose laws that attempt to address online harassment but do it carelessly, with little regard for the risks for legitimate speech. For example, recently the New York Court of Appeals struck down a cyberbullying law that made it a crime to "harass, annoy, threaten...or otherwise inflict significant emotional harm on another person," because it reached "far beyond the cyberbullying of children." After all, protected speech could very well be "annoying," but that is hardly enough reason to outlaw it.

Instead, we encourage Internet platforms to provide their users with the tools that they need to protect themselves from unwanted and abusive behavior online, including harassment or discrimination on the basis of religious, racial, ethnic, national, gender, sexual orientation. We believe that the provision of technical tools such as filters, blocklists, and reporting mechanisms, when under the control of users, can be more effective than blanket laws or policies that attempt to regulate speech, as well as being less capable of misapplication that could, in turn, infringe on the free expression rights of speakers.[5]

5 Content regulation processes

Platforms are at liberty to choose whether to moderate user-submitted content in advance, to review content that has been flagged for moderation, or to engage in proactive monitoring of content. We oppose platforms being required by law to choose any one of these models. In particular, for companies to be made subject to a general monitoring obligation would create significant obstacles to the technical and commercial feasibility of providing the kind of interactive social media platforms upon which millions of users rely.

Online publishers are not required to moderate their content in the US. They are protected under the intermediary liability provision of 47 U.S.C. § 230, which was enacted as part of the Communications Decency Act (and is sometimes called CDA 230).

Currently, most online hosting providers—including platforms like Facebook and Twitter—ban

4 See https://www.eff.org/deeplinks/2017/11/us-federal-court-rejects-global-search-order.

5 This section of our response is based on https://www.eff.org/deeplinks/2015/01/facing-challenge-online-harassment, which can be consulted for additional information.

harassment in their terms of service, but do not proactively police user behavior. Instead, they rely on community policing, or flagging, to locate and remove content or user accounts that violate their terms of service. Reports are sent to moderation teams that are often poorly supported, remotely managed, and paid considerably less than most other tech workers. Decisions about content are made quickly (sometimes within seconds), and erroneous takedowns of flagged content or accounts are fairly common.

Typically, when a user flags a piece of content, it goes into a queue to be evaluated by a content moderator, either an employee of the company, a volunteer, or—most often—a contract worker who may operate outside of the country in which the content was posted. The moderator must quickly identify whether the flagged piece of content violates the site's policies, and take some course of action. The possible types of action range widely from platform to platform—they may involve adding a filter that requires users to verify their age to view the content, removing the piece of content entirely, or shutting down the account of the user that posted it.

In some cases, users have the option to appeal the action taken against them, though this option may be limited. Generally, users receive an email or in-app notification when their content is taken down and, when applicable, the company directs users on next steps for an appeal. Often, an appeal reroutes content back through the moderation process and requires a second moderator to assess whether the original course of action was made in error.

EFF recommends that every user should have the right to due process, and that companies must provide all users with the option to appeal their takedown decisions in every case.[6] The Manila Principles provide a framework for this.

6 Bias and non-discrimination

In the US, companies generally have the legal right to choose to host, or not host, online speech at their discretion. We have spent considerable time looking at how they make those choices and have found that their practices are uneven at best, and biased at worst. Political and religious speech is regularly censored, as is nudity. In Vietnam, Facebook's reporting mechanisms have been used to silence dissidents. In Egypt, the company's "real name" policy, ostensibly aimed at protecting users from harassment, once took down the very page that helped spark the 2011 uprising. And in the United States, the policy has led to the suspension of the accounts of LGBTQ activists. Examples like these abound, making us skeptical that a heavier-handed approach by companies would improve the current state of abuse reporting mechanisms.

Examples of companies failing to take into account cultural particularities, social norms, artistic value, and other relevant interests when evaluating compliance with terms of service include the following:

- Facebook's near-blanket ban on nudity on its platform resulted in the removal of the iconic and historical Vietnam war photo of "the girl in the picture", Kim Phuc.[7]

6 https://www.eff.org/deeplinks/2018/01/private-censorship-not-best-way-fight-hate-or-defend-democracy-here-are-some

7 https://www.eff.org/deeplinks/2016/09/facebooks-nudity-ban-affects-all-kinds-users

- Twitter suspended activist Daniel Sieradski for engaging in counterspeech with hate groups.[8]
- YouTube removed WWII videos of the US Army destroying the Nuremburg swastika.[9]
- YouTube removed videos of Syria that will be eventually used in war tribunals.[10]
- Wired Magazine's analysis of Facebook removals shows its impact on Queer activists and other marginalized users.[11]

These measures are not even applied consistently across jurisdictions. For example, until this year, Microsoft Bing restricted sex-related searches throughout the entire Middle East region, including in countries with relatively liberal speech laws in which such content is lawful.[12]

As to measures that companies should adopt to prevent or redress the takedown of permissible content, please see the following section.

7 Appeals and remedies

Companies should enable users to appeal mistaken or inappropriate restrictions, takedowns or account suspensions.[13] Although some do, others do not, or only do so in certain cases. For example, Facebook users cannot appeal content takedowns during an automatic ban period, and cannot appeal decisions about individual posts, photos, or videos. Instagram allows for appeals in the case of an account suspension, but not when content is removed.

In the case of content uploaded to YouTube, a match against YouTube's ContentID, which is used for copyright enforcement, can trigger that content to be automatically restricted. In such a case an eligible user (a broad category that appears to include verified users "in good copyright standing") can dispute the ContentID match, and where the rightsholder has rejected that dispute, can file an appeal. In the case of an appeal, the copyright holder must either release the claim or file a formal takedown under the U.S. Digital Millennium Copyright Act (DMCA).

For more information on major platforms' appeals and remedies, we provide a guide to appealing content decisions at Onlinecensorship.org.[14]

8 Automation and content moderation

Increasingly, companies are seeking ways to increase the efficiency and speed of content moderation. Content moderation by human moderators is a resource-intensive process, and results in significant levels of burnout among moderators employed to engage in clickwork that involves watching often violent and abusive content.

As a result, companies are turning to algorithmic means to automate the moderation process where

8 http://forward.com/fast-forward/374276/antifas-most-prominent-jew-booted-from-twitter/
9 https://boingboing.net/2017/08/14/film-of-u-s-army-destroying-n.html
10 http://www.middleeasteye.net/news/youtube-criticised-after-middle-east-video-taken-down-over-extremist-content-1244893230
11 https://www.wired.com/story/facebooks-hate-speech-policies-censor-marginalized-users
12 https://www.eff.org/deeplinks/2017/07/microsoft-bing-reverses-sex-related-censorship-middle-east
13 Manila Principles on Intermediary Liability, paras 5(b) and 5(c).
14 https://onlinecensorship.org/resources/how-to-appeal

possible. This may include the use of PhotoDNA, a technology that computes hash values for images or audio files, reducing them to a digital signature that can be used to identify identical images or files. PhotoDNA is used in Project Vic, an initiative of the National Center for Missing and Exploited Children, to check retrieved images of child sexual abuse imagery to help law enforcement officials identify and locate missing children. Another approach involves using systems like YouTube's ContentID, which, similar to PhotoDNA technology, automatically removes content it identifies as copyrighted and uploaded by a user to YouTube's systems. YouTube has also been piloting the use of artificial intelligence (AI) to flag extremist content.

There are a number of problems with automatic content filtering. For example, in December 2017 a user who uploaded a public domain recording of audio from NASA triggered a false copyright match against music recording of Pink Floyd that also used the same public domain NASA clip,[15] and in another case a ContentID match was triggered by an original live recording of birdsong.[16]

In addition, companies including Facebook and Twitter have imposed automatic bans for certain types of content and for multiple offenses by users. These bans range from 12 hours to 30 days, depending on the offense, and cause the user to be automatically locked out of their account, essentially instituting a "cool-down period" before they are allowed back on the platform. Automatic bans cannot be appealed by the user.

9 Transparency

Users should be notified about content restrictions, takedowns, and account suspensions, including the reasons therefor, the policies pursuant to which these take place, and the available avenues for redress and appeal.[17] Platforms should also be transparent to the general public about their content restriction actions. This should include actions taken on government requests, court orders, private complainant requests, and enforcement of terms of service.[18] Currently, many platforms only provide information on actions taken on government requests. These are usually claims of illegal content, and the numbers usually do not include government requests for takedowns based on terms of service.

It is also possible for platforms to identify the removal of information by inserting a note at the location from which the information was removed.[19] For example, when attempting to access a search result that has been removed by Google under the DMCA, the user may see a message such as "In response to a complaint we received under the US Digital Millennium Copyright Act, we have removed 1 result(s) from this page. If you wish, you may read the DMCA complaint that caused the removal(s)". The archival of such DMCA complaints is undertaken in partnership with the independent transparency website, Lumen.[20]

Regrettably, the same transparency is not possible for search results removed under the European Union's "right to be forgotten". Instead, in response to any search request for what appears to be an individual's name, Google prints a blanket statement that results may have been removed at the

15 https://arstechnica.com/tech-policy/2017/12/facebook-sends-ars-takedown-notice-from-pink-floyd-over-nasa-audio/
16 https://yro.slashdot.org/story/12/02/26/2141246/youtube-identifies-birdsong-as-copyrighted-music
17 Manila Principles on Intermediary Liability, paras 5(a) and 6(c).
18 Manila Principles on Intermediary Liability, para 6(e).
19 Manila Principles on Intermediary Liability, para 6(f).
20 https://lumendatabase.org/

bottom of the search page.

Most companies tell us almost nothing about when they remove content or restrict users' accounts for violating their rules. Through their terms of service and user agreements, companies set their own rules for what types of content or activities are prohibited on their services and platforms, and have their own internal systems and processes for enforcing these rules. Companies need to disclose more information about their enforcement processes and the volume and nature of content being removed. Furthermore, best practice is for companies to inform users about the specific rule that they've violated and the specific terms of their content takedown.[21]

10 Examples

Examples of content regulation that raise freedom of expression concerns for EFF are maintained on our project website Onlinecensorship.org. Some examples drawn from that site include the following:

- Facebook banned a bra company from advertising its guide of breast sizes and shapes, even though the ad only showed a drawer full of bras.[22]
- The "Innocence of Muslims"video sparked anger in some countries, so YouTube independently blocked it in Egypt and Libya.[23]
- Facebook shut down the account of HappyAddis, a prominent LGBT activist, for failing to abide by the Real Name Policy. Same-sex relations are criminal in Ethiopia.[24]
- Popular rapper Talib Kweli used contextualized imagery to make a statement about racism, but Instagram took it down as "hate speech".[25]
- Twitter wants you to abstain...from advertising. It banned the National Campaign to Prevent Teen and Unplanned Pregnancy's sex-positive health messages.[26]
- Artist Rupi Kaur wanted to demonstrate society's uneasiness about menstruation. She didn't expect Instagram to prove her point by censoring the photos.[27]
- Former Facebook employee and trans woman Zoë Cat had to show proof of her name under the platform's Real Name Policy.[28]
- Facebook removed a video of the self-immolation of a Tibetan monk. 136 Tibetans have

21 An Onlinecensorship.org research paper, conducted with the Queensland University of Technology, is currently under preparation and will be released in 2018. See: https://digitalsocialcontract.net/new-project-creating-better-standards-for-transparency-in-how-platforms-moderate-content-fc7ba1b44fdb.
22 http://kernelmag.dailydot.com/issue-sections/features-issue-sections/12796/facebook-nudity-breasts-advertising/
23 http://www.washingtonpost.com/business/economy/googles-restricting-of-anti-muslim-video-shows-role-of-web-firms-as-free-speech-arbiters/2012/09/14/ec0f8ce0-fe9b-11e1-8adc-499661afe377_story.html
24 http://time.com/money/3954390/ethiopian-lgbt-activist-banned-facebook-real-name/?
25 http://mashable.com/2015/07/01/instagram-talib-kweli-hate-speech/
26 http://www.theatlantic.com/health/archive/2015/03/when-social-media-censors-sex-education/385576/
27 http://www.huffingtonpost.co.uk/2015/07/23/rupi-kaur-instagram-censorship-artist-period-milk-and-honey_n_7836108.html
28 http://www.slate.com/articles/technology/future_tense/2015/07/facebook_s_authentic_name_policy_ensnares_zo_cat_a_trans_woman_who_tried.html

self-immolated since 2009 to protest Chinese restrictions on their culture and religion.[29]

- The 2011 Egyptian uprising was first announced on a page that Facebook had deleted a year before because one of the administrators was using a pseudonym.[30]
- In 2011, Flickr banned prominent Egyptian activist Hossam Hamalawy for posting photos retrieved from a raid on state security offices.[31]
- After the Charlie Hebdo shootings, Mark Zuckerberg spoke out for free expression. Two weeks later, Facebook began censoring some images of the Prophet.[32]
- Facebook censored a photograph of a Brazilian indigenous couple from 1909 because in the picture, the female appeared bare breasted.[33]

29 http://sinosphere.blogs.nytimes.com/2014/12/27/facebook-deletes-post-on-tibetan-monks-self-immolation/?_r=0
30 http://www.thedailybeast.com/articles/2011/02/24/middle-east-uprising-facebooks-back-channel-diplomacy.html
31 http://www.nytimes.com/2011/03/28/business/media/28social.html
32 https://www.washingtonpost.com/news/the-intersect/wp/2015/01/27/two-weeks-after-zuckerberg-said-je-suis-charlie-facebook-begins-censoring-images-of-prophet-muhammad/
33 http://m.abc.com.py/internacionales/facebook-restituye-foto-de-indigena-desnuda-tras-queja-de-gobierno-de-brasil-1357854.html