UNFRIENDING CENSORSHIP

Insights from four months of crowdsourced data on social media censorship

By Jessica Anderson, Matthew Stender,
Sarah Myers West, and Jillian C. York

ELECTRONIC FRONTIER FOUNDATION

VISUALIZINGIMPACT

This document is the first report released by Onlinecensorship.org, drawing from four months of user-sourced data gathered between November 2015 and March 2016 and covering six prominent social media platforms. Onlinecensorship.org seeks to encourage social media companies to operate with greater transparency and accountability toward their users as they make decisions that regulate speech.

Onlinecensorship.org is a collaboration between the Electronic Frontier Foundation (EFF) and Visualizing Impact (VI). EFF is an international organization with nearly 30,000 members worldwide from over 100 countries, dedicated to the protection of everyone's privacy, freedom of expression, and association online. VI is an interdisciplinary collective that applies data journalism, technology, and design to social issues.

Follow the work of Onlinecensorship.org:
@censored
https://www.facebook.com/Onlinecensorship

Contact us about this report:
jillian@eff.org | jessica@visualizingimpact.org

TABLE OF CONTENTS

1. INTRODUCTION

Onlinecensorship.org was founded to fill a gap in public knowledge about how social media companies moderate content. As sites like Facebook, Twitter, and YouTube play an increasingly prominent role as platforms for digital speech, the companies that run them are serving as gatekeepers, determining what constitutes "acceptable" content for public consumption. As self-ordained content moderators, they face increasingly thorny issues; deciding what constitutes hate speech, harassment, and terrorism is challenging, particularly across many different cultures, languages, and social circumstances.

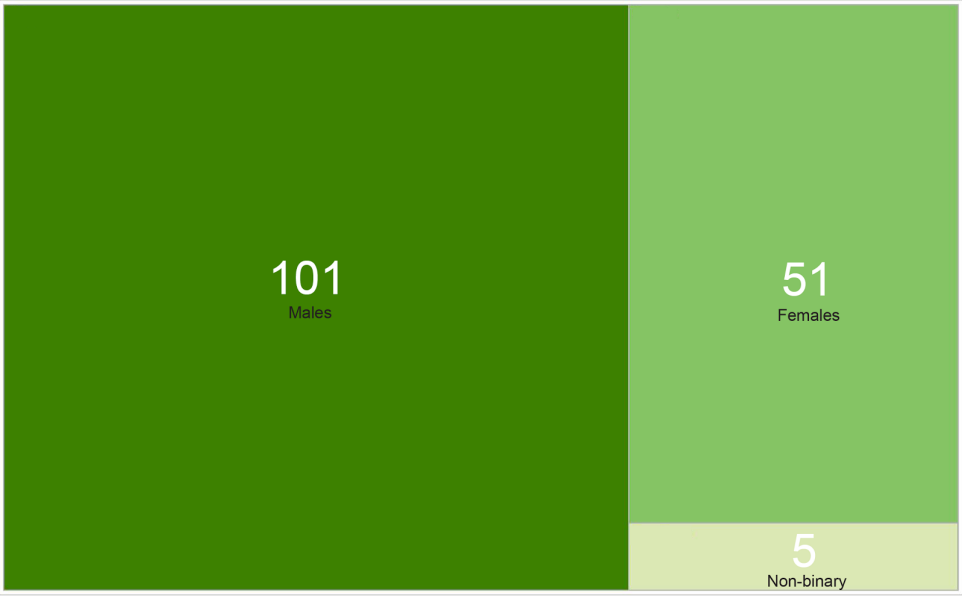
Some users treat popular social media platforms as the proverbial "public sphere," a space collectively owned and open to activism and debate. Others treat their social media as a very personal space that is theirs to cultivate according to their individual identity. Both attitudes can come into conflict with the actions of corporations, which have a fiduciary responsibility to their shareholders, not to their users. These platforms have amassed staggering amounts of user metadata from across the world, assert rights over users' content, and employ their own rules and systems of governance that control—and in some cases, censor—it.

In the United States (where all of the companies covered in this report are headquartered), social media companies generally reserve the right to determine what content they will host, and they do not consider their policies to constitute censorship. We challenge this assertion, and examine how their policies (and the enforcement thereof) may have a chilling effect on freedom of expression. Although "censorship" traditionally refers to "official" regulation of expression, it is our belief that these sites have become so large in scale that the impact of their content moderation decisions on the world is outsized. Currently, there is a lack of transparency surrounding content moderation decisions, as well as the processes through which users can appeal to restore their content when it is removed. In order to encourage companies to operate with greater transparency and accountability toward their users, we have collected reports from users. Our data works to shine a light on what content is taken down, why companies make certain decisions about content, and how content takedowns affect communities of users around the world.

This is our inaugural report on the submissions we've received through the Onlinecensorship.org platform. We asked users to send us reports when they had their content or accounts taken down on six social media platforms: Facebook, Flickr, Google+, Instagram, Twitter, and YouTube. Through a questionnaire on the Onlinecensorship.org website, they answered a series of questions about how the content was taken down and the impact this had on their lives. We have aggregated and analyzed the collected data across geography, platform, content type, and issue areas to highlight trends in social media censorship. All the information presented here is anonymized, with the exception of case study examples we obtained with prior approval by the user. The data we present comes directly from users. While we introduced some measures to help us verify reports (such as giving respondents the opportunity to send us screenshots that support their claims), we did not work with the companies to obtain this data and thus cannot claim it is representative of all content takedowns or user experiences. Instead, it shows how a subset of the millions of social media users feel about how their content takedowns were handled, and the impact it has had on their lives.

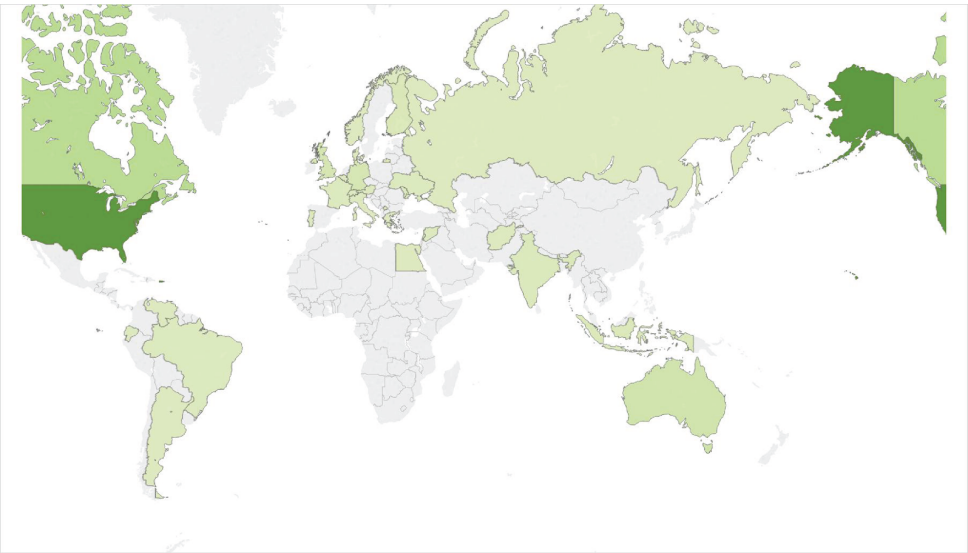
2. OVERVIEW OF DATA

RESPONDENTS BY AGE GROUP



RESPONDENTS BY GENDER

PRIMARY LANGUAGE OF POSTED CONTENT

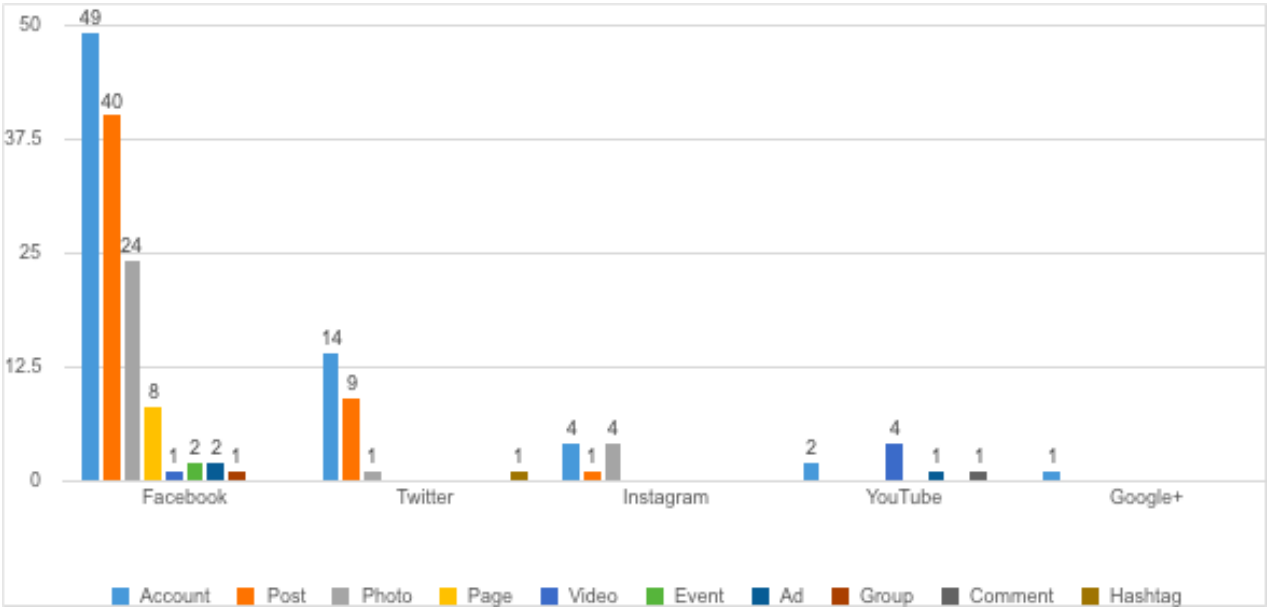


93 US, 22 Canada, 5 UK, 3 India, 2 France, 2 Ecuador, 2 Ukraine, 1 Brazil, 1 Venezuela, 1 Argentina, 1 Portugal,1 Denmark, 1 Switzerland, 1 Austria, 1 Italy, 1 Greece, 1 Moldova, 1 Norway, 1 Finland, 1 Egypt, 1 Syria, 1 Lebanon, 1 Russia, 1 Afghanistan, 1 Indonesia, 1 Australia

NUMBER OF REPORTS BY COUNTRY

Onlinecensorship.org launched officially in November 2015. This report describes information collected over a period of four months, between November 19, 2015 and March 10, 2016. During that period, we received 186 reports. After excluding incomplete or erroneous reports (for example, reports on platforms not covered by Onlinecensorship.org), we were left with 161 reports. Of the users who provided their age and gender, just over half of the respondents fell between the ages of 25-44 and identified as male; just under half of the respondents identified as either female or non-binary.

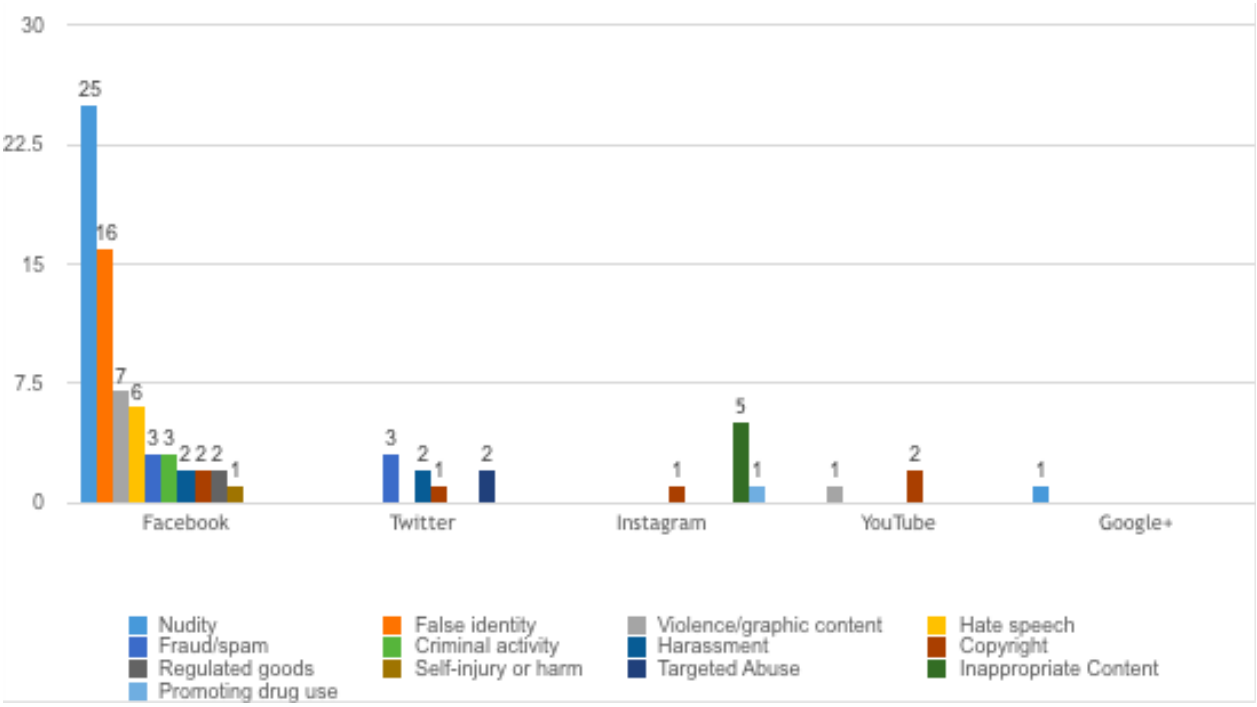
User submissions spanned eleven different languages, though there was a heavy skew toward English, likely influenced in part by the fact that the Onlinecensorship.org site is currently in English only (Arabic and Spanish will soon be added). Unsurprisingly then, there was a strong corresponding geographic skew toward the United States, with Canada, Australia, and the United Kingdom the next most-represented countries.



TAKEDOWNS BY PLATFORM AND TYPE

Facebook takedowns were the subject of a notable majority of submissions, followed by Twitter. We received no submissions from Flickr users. Account suspensions, the most stringent form of moderation, were the most frequent type of content takedown in our dataset. We don't yet have enough information on content takedowns to judge if this finding is representative of what is happening across all content takedowns, as most companies do not provide this information. It is possible that users who experienced full account shutdowns were more motivated to report their takedown experience on Onlinecensorship.org. The second most common form of moderation was the removal of a post, both for Facebook and for Twitter.

"Account suspensions, the most stringent form of moderation, were the most frequent type of content takedown in our dataset."



REASON FOR CONTENT TAKEDOWN BY PLATFORM

Based on what we have seen so far, several illustrative trends are emerging regarding the reason given to the user for the content takedown:

- The majority of reported content takedowns for Facebook were due to either nudity or use of a false identity. Given the prominence of these two issues in the data, we analyze these submissions in greater detail in the next section.
- Instagram users tended to report "inappropriate content" as the reason their content was taken down. Several of these users said they were confused by the vagueness of this reason, and didn't understand why their work was removed.
- Takedowns on Twitter tended to be linked to targeted abuse, harassment, or fraud/spam, which may reflect the company's recent focus on combating abusive behavior on the platform.
- Nearly half of our copyright-related takedown submissions came from YouTube. While these submissions are suggestive of user perceptions that copyright flags are used for censorship, a more complete picture of copyright takedowns may be found in the data collected by Lumen (formerly Chilling Effects), a database which collects and analyzes these kinds of requests.[1]

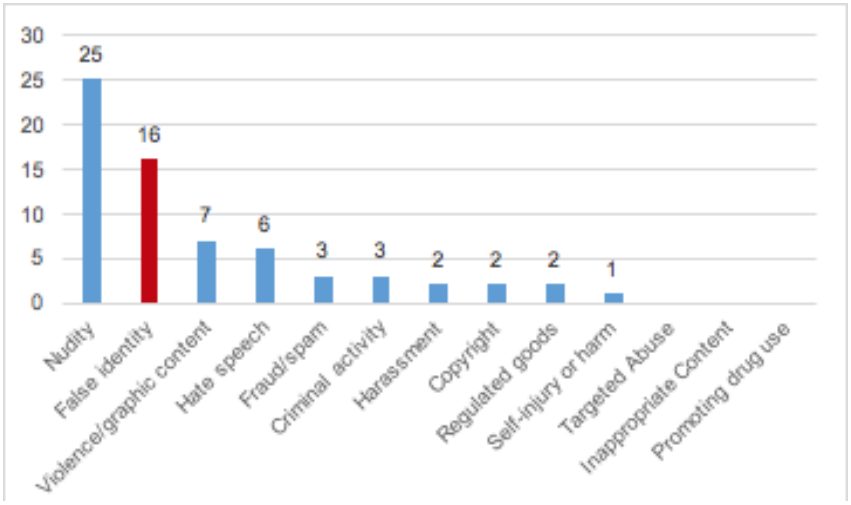
[1] "Lumen Database," https://lumendatabase.org/.

3. IN FOCUS: TAKEDOWN TRENDS

Facebook's "Real Names" Policy

Of 119 reports we received from Facebook users, sixteen (13%) reported they had been asked to prove their identity to Facebook under its name policy. Unlike the other platforms in our survey, Facebook employs a strict policy requiring individuals to use a name "their friends or family know them by" on Facebook. This policy has come under scrutiny by the media and rights groups,[2] particularly for the impact that it can have on marginalized communities.[3] Although Facebook made moderate changes to its policy in 2015 following the mass suspension of drag performers' profiles, users continue to be suspended regularly for various violations.[4]

In the media: Former Facebook employee and trans woman Zoë Cat had to show proof of her name to Facebook.[5]



FALSE IDENTITY REPORTS

[2] Jillian York, "A Case for Pseudonyms," *Deeplinks*, July 29, 2011, https://www.eff.org/deeplinks/2011/07/case-pseudonyms.
[3] Dia Kayyali and Jillian York, "Facebook's 'Real Name' Policy Can Cause Real-World Harm for the LGBTQ Community," *Deeplinks*, September 16, 2014, https://www.eff.org/deeplinks/2014/09/facebooks-real-name-policy-can-cause-real-world-harm-lgbtq-community.
[4] Wafa Ben Hassine and Eva Galperin, "Changes to Facebook's "Real Names" Policy Still Don't Fix the Problem," *Deeplinks*, December 18, 2015, https://www.eff.org/deeplinks/2015/12/changes-facebooks-real-names-policy-still-dont-fix-problem.
[5] "What's in Real Name," accessed March 23, 2016, http://www.slate.com/articles/technology/future_tense/2015/07/facebook_s_authentic_name_policy_ensnares_zo_cat_a_trans_woman_who_tried.html.

Name-related submissions at a glance:

- In sixteen cases (13% of total Facebook cases), accounts were suspended on the basis of users' names
- In ten cases (63%), users chose to appeal Facebook's decision
- 40% of the time, users who appealed said that Facebook never responded
- 50% of the users who didn't appeal said that they did not know they could appeal Facebook's decision

Real name case types:

- Legal name mistakenly (or intentionally) flagged as false
- Non-legal name central to a user's identity, or that a user's friends or family know them by, but that Facebook does not accept as "real"
- Non-legal name not central to a user's identity and/or not necessarily known by user's friends or family. Users with these names sometimes, but not always, provide context for their name choice that includes concerns for online safety.

While we cannot always verify which of the above types our cases fall into, users' experiences across all types raised questions about why and how a particular account comes under scrutiny among the millions of false accounts on Facebook.

Emerging themes in real name cases:

- Possible abuse of reporting mechanisms: In at least four cases (25%), users perceived a direct link between their name being flagged as false and their recent activities, such as authoring a controversial post or participating in a heated debate. In these cases, reporting a user for a false name may be a proxy for dealing with other issues such as "offensive" content (as defined by the user), trolling, or harassment.

"Reporting a user for a false name may be a proxy for dealing with other issues such as "offensive" content (as defined by the user), trolling, or harassment."

- Possible bias against users from developing markets: Facebook notes in its 2014 annual report: "We believe the percentage of accounts that are duplicate or false is meaningfully lower in developed markets such as the United States or United Kingdom and higher in developing markets." We received reports from two users that reported being flagged for having a name (or letter in their name) that is unfamiliar to English speakers. They suspect that users from developing markets may be subject to bias or higher scrutiny in name assessments.

- Strong preference for legal names: Facebook made changes to its "real name" policy in December 2015 that allow users to go by their "authentic name,"[6] described as the name "their friends or family know them by." Two users noted they were using non-legal names or nicknames central to their identity and that most of their friends and family know them by, but had their accounts removed nevertheless.
- Lack of transparency linked to user distrust: Several users reported that the process for recovering their accounts or appealing Facebook's decision was confusing, and that Facebook's decision-making process was impenetrable and impersonal. Though some users attempted to comply with Facebook's request for ID, others refused the procedure, expressing distrust of Facebook's motives and questioning their respect for user privacy.
- Accounts in limbo: Users reported that their suspended accounts were neither closed completely nor accessible, leaving them in limbo. Users were frustrated that they could not recover data from Facebook. They also continued to receive notifications from other users with no ability to respond, which they felt hurt them socially.

Cases:

1: *"It [the real name process] is opaque. Completely a cover for saying 'no' with no actual human involvement."*

The user posted a piece of controversial content criticizing the movement for transgender rights. The same day, Facebook contacted him to request proof of identity. The user suspected that he had been reported by someone who disagreed with his politics. He did not comply with Facebook's request for identification, and after seven days, his account was suspended. He says it has remained in a state of review for several months. It should be noted that we received similar reports across the political spectrum.

2: *"I refuse to provide private information...to an organisation that has complete contempt for privacy."*

The user describes suspecting "a sustained campaign from others to kill my account." Additionally, he was uncomfortable with providing ID to Facebook. Although Facebook allows users to cover parts of their IDs that are not relevant to the verification of their name, some users may be unaware of this or unsatisfied with the level of security provided by that concession.

[6] Amanda Holpuch, "Facebook adjusts controversial 'real name' policy in wake of criticism," the *Guardian*, December 15, 2015, http://www.theguardian.com/us-news/2015/dec/15/facebook-change-controversial-real-name-policy.

3: *"I provided them with everything they said they wanted and they still said no...All my photos, hundreds of thousands of messages, friends from around the world, are still lost. In other people's chat logs, all my messages have been deleted and marked as 'spam or abusive.'"*

The user, who is from a non-English speaking country, has a name likely to be unfamiliar to English speakers. She complied with Facebook's initial request for ID. The ID type she submitted apparently qualified, but Facebook suspected that it was not valid and requested two further forms of identification, which she provided. Still, Facebook did not restore her account, and she says they offered no final justification for this decision beyond skepticism of the submitted IDs.

4: *" When they pull the real name game on users, it is always under duress and never when one is signing up for an account."*

This user speculates that Facebook may have initially flagged her name due to a special character it contained, the umlaut. Although she did not fully comply with the name verification process, her main concern centered on how Facebook raises issues with users' names after they have consolidated their community and memories on the platform, rather than at signup. In doing so, she feels that Facebook creates a coercive environment for decision-making. She also referenced the experience of a friend who was flagged for using a religious name, which Facebook does not permit. He modified his account to comply and reappeared under "a name he was not known by in general or with friends/associates."

Nudity

The banning of nude imagery by a number of social media platforms has been a hot topic for some time. The removal of certain types of nude imagery—particularly breastfeeding and post-mastectomy photos—has led some users to organize around their concerns, leading to significant media coverage. The Free The Nipple campaign,[7] which advocates for female toplessness to be treated equally to male toplessness, appears to have had some influence on company policies over the past few years, leading, for example, to Facebook clarifying its policies in March 2015.[8]

[7] "Wikipedia: Free the Nipple (campaign)," accessed March 23, 2016, https://en.wikipedia.org/wiki/Free_the_Nipple_(campaign).
[8] Stuart Dredge, "Facebook clarifies policy on nudity, hate speech and other community standards," the *Guardian*, March 16, 2015, http://www.theguardian.com/technology/2015/mar/16/facebook-policy-nudity-hate-speech-standards.

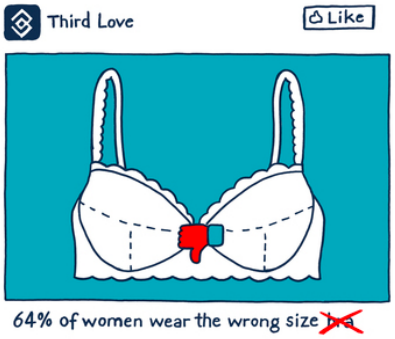
Instagram's current Community Guidelines ban "most nudity" for "a variety of reasons" which are not explicated. The Guidelines note specifically that the ban "includes some photos of female nipples" but that "photos of post-mastectomy scarring and women actively breastfeeding are allowed" and "nudity in photos of paintings and sculptures is OK, too."

Similarly, Facebook's Community Standards specifically state that the company "always allow[s] photos of women actively engaged in breastfeeding or showing breasts with post-mastectomy scarring" as well as "photographs of paintings, sculptures, and other art that depicts nude figures." Digitally-created content may not contain nudity or sexual activity "unless the content is posted for educational, humorous, or satirical purposes." Unlike Instagram, Facebook explicitly states that nudity is restricted "because some audiences within our global community may be sensitive to this type of content - particularly because of their cultural background or age."

Google+ bans sexually explicit content, but makes broader exceptions for "naturalistic and documentary depictions of nudity (such as an image of a breastfeeding infant), as well as depictions of nudity that serve a clear educational, scientific, or artistic purpose."

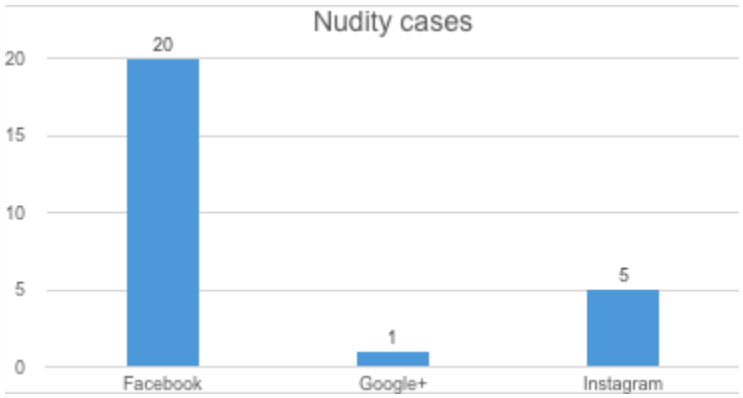
Flickr and Twitter allow nude imagery, though both place restrictions on where such content may appear on their sites.

In the Media: Twitter banned the National Campaign to Prevent Teen and Unplanned Pregnancy from posting a health-related, sex-positive message.[9]



In the Media: Facebook banned ThirdLove, a bra company, from advertising its illustrated reference guide of various breast sizes and shapes, even though the ad only showed a drawer full of bras.[10]

[9] "When Social-Media Companies Censor Sex Education," accessed March 23, 2016, http://www.theatlantic.com/health/archive/2015/03/when-social-media-censors-sex-education/385576/
[10] "Why Facebook Still Has a Problem with Breasts," accessed March 23, 2016, http://kernelmag.dailydot.com/issue-sections/features-issue-sections/12796/facebook-nudity-breasts-advertising/



NUDITY REPORTS AS % OF TOTAL REPORTS PER PLATFORM

Nudity-related submissions at a glance:

- Twenty-six cases of content takedowns and/or temporary bans occurred due to content containing nudity. One of those cases was on Google+ and five on Instagram; the rest were on Facebook.
- Of those twenty-six cases, four (15%) reportedly contained male nudity, while seven (27%) reportedly contained female nudity. In the remaining 15 cases, the type of nudity depicted was not reported.
- In fourteen cases (36%), users chose to appeal Facebook's decision. Content was restored in one case.
- One case involved an image of a mother breastfeeding and resulted in both the takedown of the user's photo and suspension of their account. An appeal by the user was unsuccessful in restoring either.

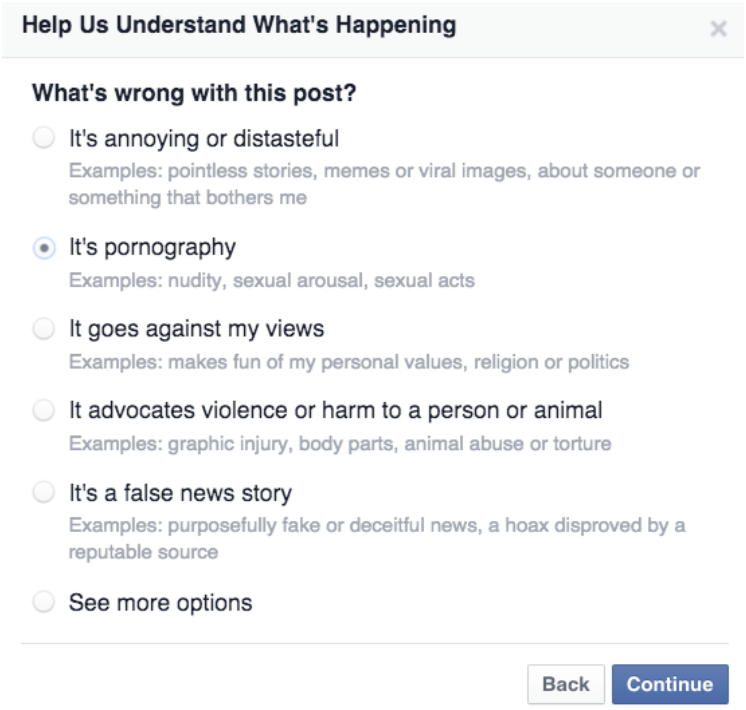
Nudity case types:

- Users who knowingly and/or defiantly violated the Community Standards
- Users whose breastfeeding images were taken down despite exceptions in the Community Standards for such content
- Users whose nude artwork was taken down despite exceptions for such content

Emerging themes in nudity cases:

- Content for which the Facebook Community Standards have carved-out exceptions is still being removed. The current Community Standards allow for both breastfeeding images and photographs of artwork containing nudity. Nonetheless, several users reported that their nude artwork or breastfeeding images were taken down and that subsequent appeals to restore the content were unsuccessful.
- Sexual health information is being censored. Several reports we received were in relation to content that we believe serves an educational purpose.

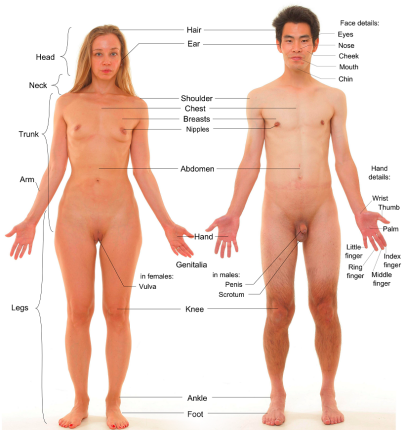
- Users are subject to increasingly long temporary bans for repeat offenses. While some users reported intentionally posting nude images to challenge platform policies, the punishment for doing so is severe. Users may be banned from the platform for 24 hours or longer for repeat offenses; one user reported being banned for a month for posting his nude photography.
- Facebook treats pornography and nudity as one and the same. Facebook's reporting mechanism for images allows users only to report "pornography," an umbrella term under which the company includes nudity as well as "sexual arousal" and "sexual acts." Users report that the images taken down were single-subject and not sexual in nature.



Cases

1. "The content was the Wikipedia photo of human anatomy showing a man and a woman in full frontal nudity. It is against Facebook's guidelines about nudity."

The user posted an image from Wikimedia Commons (at right) depicting a nude man and woman side by side, with body parts anatomically labeled.[11] The image is clearly for educational purposes, but was removed for violating the Community Standards. The user did not appeal, because he realized that he had violated the rules.



[11] "Wikimedia Commons: Anterior view of human female and male, with labels," accessed March 23, 2016, https://commons.wikimedia.org/wiki/File:Anterior_view_of_human_female_and_male,_with_labels_2.png.

2. "I find that open source and private censorship of Facebook is very important and quite alarming."

The user paid for an advertisement for his non-partisan political advocacy group that contained a digitally-created image of Atlas, from the novel *Atlas Shrugged*, handing the world from his shoulders to a group of people. Although the image was digitally-created and contained no genitalia, the ad was rejected and the user received a message stating: "We recommend using content that focuses on your service or product rather than nudity." The user believes that an "*Atlas Shrugged* supporter marked it as [an] offensive nude picture because they didn't like the message."

3. "I had thousands of followers on that page and I lost all the time and effort I put into compiling all the wonderful fine art."

The user operated a page that featured paintings of nude women, including paintings by famous artists such as Picasso, Renoir, and Dali, as well as some by contemporary artists. This user was aware that the rules allowed an exception for nude paintings. After a user or users reported the page, Facebook deleted it. The administrator reports that, although he appealed and Facebook responded within two days, the appeal was unsuccessful and a reason for the takedown was not provided.

4. "There is no real 'appeal' process with Facebook. You can click on their ridiculous series of 'emojis' and leave a Comment, but that's it. You might as well be speaking into a void."

The user, an award-winning photographer whose photographs have been exhibited in world-class museums, was banned several times, for 30 days each, for posting fine art nude photographs. Although the user appealed, the content was not restored. The user states: "I regard my fine art photographs to be every bit as much art as any representation of a drawing, painting or sculpture" and notes that in his jurisdiction, female toplessness is legal in public.

4. IMPACT ON USERS

The manner in which social media companies implement their regulations can seriously impact the lives of users. While content takedowns might be brushed aside as the companies' prerogative, several of the companies we've studied are no longer simply taking down content, but imposing temporary and longer-term bans on users who violate the rules. These bans are often surprising to users who may not realize the consequences of violating the rules or that they've violated the rules at all.

During the course of our research, an Onlinecensorship.org team member posted an image to Facebook from a breast cancer awareness campaign that depicted a seated woman exposing one breast.[12] When the team member tried accessing their account the next morning, they discovered that they had been logged out automatically. After logging in, they were taken through a "checkpoint" (Facebook's terminology) where they had to acknowledge they had violated the Community Standards, and were then shown their other recent posted photos and asked to remove any that might contain nudity. Following that, they were subject to a 24-hour ban on using the service. This type of ban was reflective of the experiences of several other users.

During the temporary ban, the user was prevented from posting or "liking" on the main site, although they found that they could "like" posts when using Facebook's mobile web version. They could not change their personal information or send messages, sign up for or into external services like Spotify or Tinder, or leave comments on websites like Huffington Post that use Facebook's commenting system. Most disconcerting, they could not administer Pages, several of which are work-related. This restriction included the inability to remove or add other administrators.

It is also telling which services remained available during the ban. The user could buy advertisements and use Airbnb through Facebook's login. The user could change the pronunciation of their name (a minor feature on users' "about" pages), receive messages, report other users for violating the Community Standards, and block other users.

How has this takedown impacted you?

"I feel that my voice and ideas are no longer getting out to the people on my page, especially...regarding my blog or activist information. Until recently, I had so little feedback in comparison to the past...I feel this infringes on my freedom of expression, especially due to [t]he fact that this expression highlights ideas that aren't so mainstream and are helpful to the cause of human empowerment. Please help me find a solution."

"Apparently, Facebook has branded me a troublemaker, because I am often punished for posting stuff that many others are posting without any punishment or warning...I worry that if I am ever dying & attempt to reach out to my friends, Facebook will block me from contacting them, which frightens me & causes me considerable angst."

"I am disabled and very sick...I have had my tagging ability taken away three times and my ability to post taken away once. If I cannot tag, I cannot guarantee that I am alerting anyone! If I can't post, I am really in trouble."

"I am a disabled vet, pretty much the only way I have to communicate with my family and friends is through Facebook."

[12] "Pink Ribbon Germany: Check it before it's removed," accessed March 23, 2016, http://www.checkitbeforeitsremoved.com/en/#fototeilen.

It is unclear to us how frequently such bans are enacted or what prompts them. Repeat offenders appear to be subject to bans lasting as long as 30 days, even for seemingly innocuous behavior such as posting artful nudity.

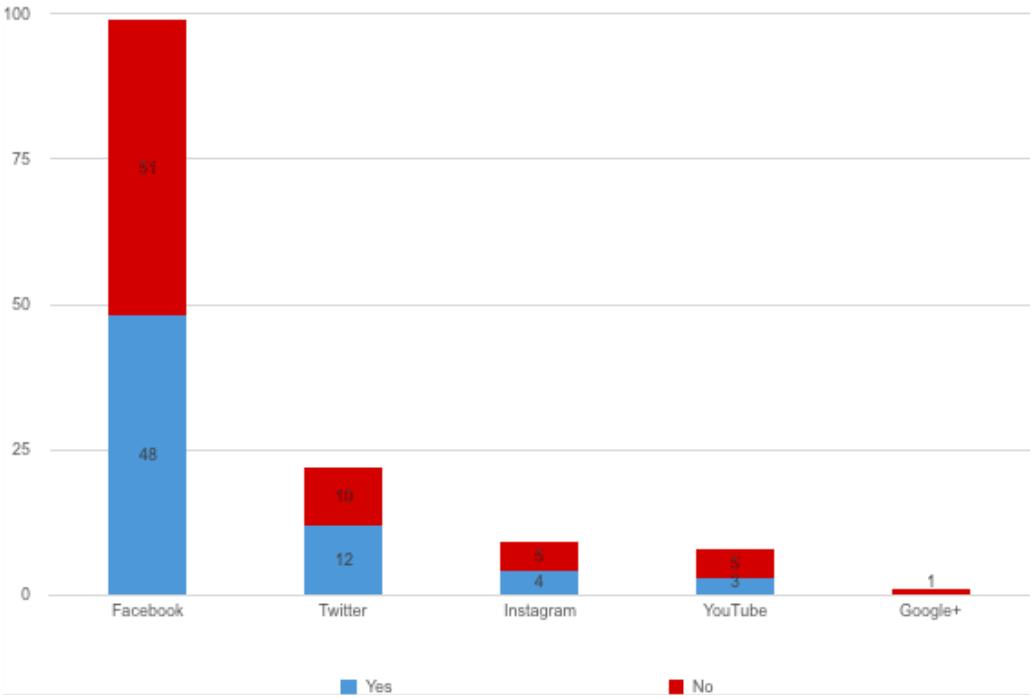
We are aware that other sites enact similar bans to those imposed by Facebook, but have not yet received any user accounts of what those bans entail.

Account deactivations, such as those imposed by Facebook for "real name" policy violations by those who cannot or refuse to submit identification, can have an even more serious impact on users' lives. Facebook is widely used for logging in to sites around the web; users whose Facebook accounts have been deactivated may lose access to countless other sites and services. As a result of how central Facebook has become to many individuals' lives, users who are banned from the service may lose access to their entire network; we received reports from a number of users who, subject to temporary bans or account deactivation, complained that they had lost contact with family or friends. Twitter users who have spent years building up a following lamented in their reports having to start over from scratch when creating a new account.

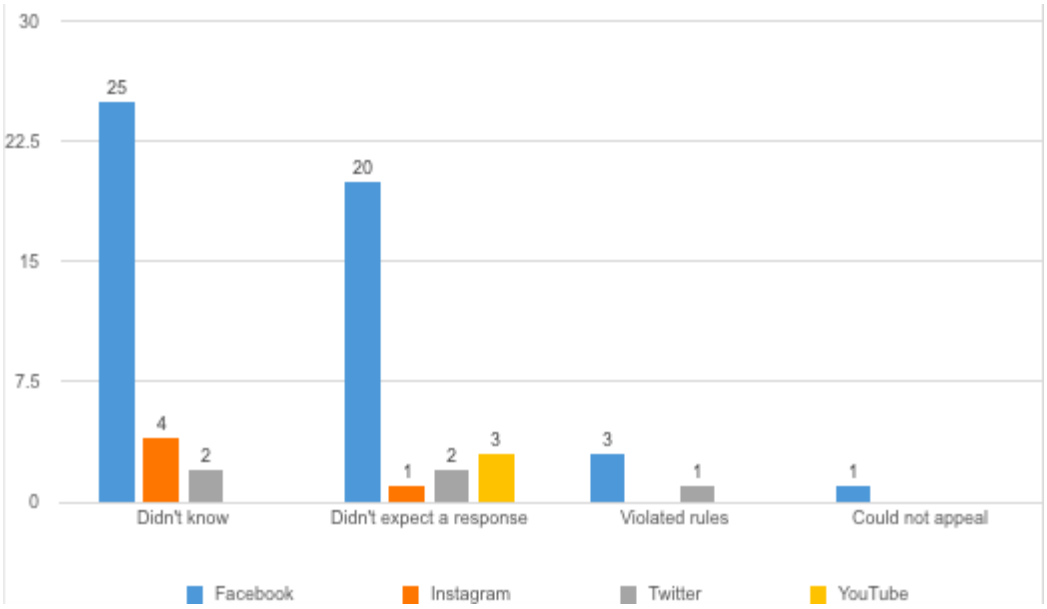
Our observation is that the impact of content takedowns and account deactivations on users varies. Some users experience only annoyance, while others (particularly those living far from their home or their chosen communities) may feel a greater impact.

Appeals

Many of the social media platforms included in Onlinecensorship.org's survey offer some mode of recourse to users whose content has been removed. However, the extent of that recourse varies widely from platform to platform: at the broadest level, Twitter says it allows users to appeal any form of action they take, while more narrowly, Instagram and Facebook only allow an appeal for the removal of a profile (or Page, in the case of Facebook).



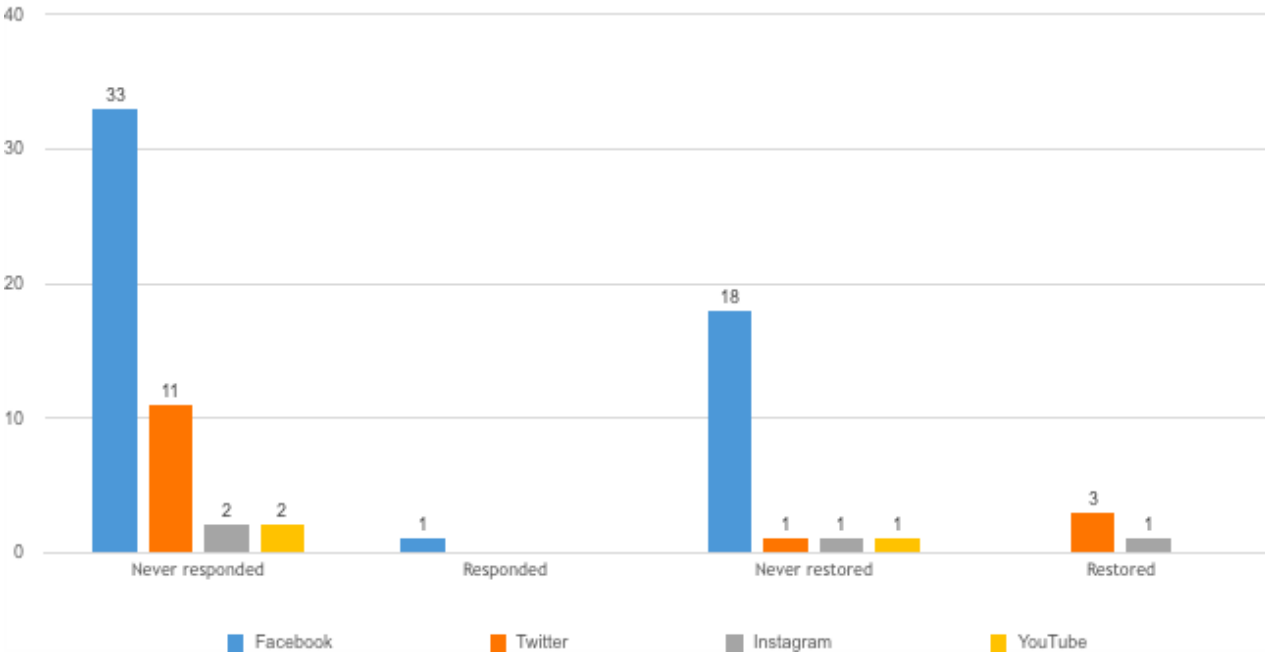
DID YOU APPEAL?



REASON FOR NOT APPEALING

Our resource guide to appealing content removals reflects the challenge of effectively conducting this process; many of the platforms have dedicated teams that cover different time zones and languages in order to accommodate users' appeals. Despite these efforts, reports collected so far suggest that the appeals process is confusing, difficult to navigate, and rarely results in success for the user. In only four cases were users able to have their account or content successfully restored, while in almost fifty they reported not even receiving a response. Many said they did not know that they were able to appeal, or that they did not know how. Others reported they did not expect a response even if they tried.

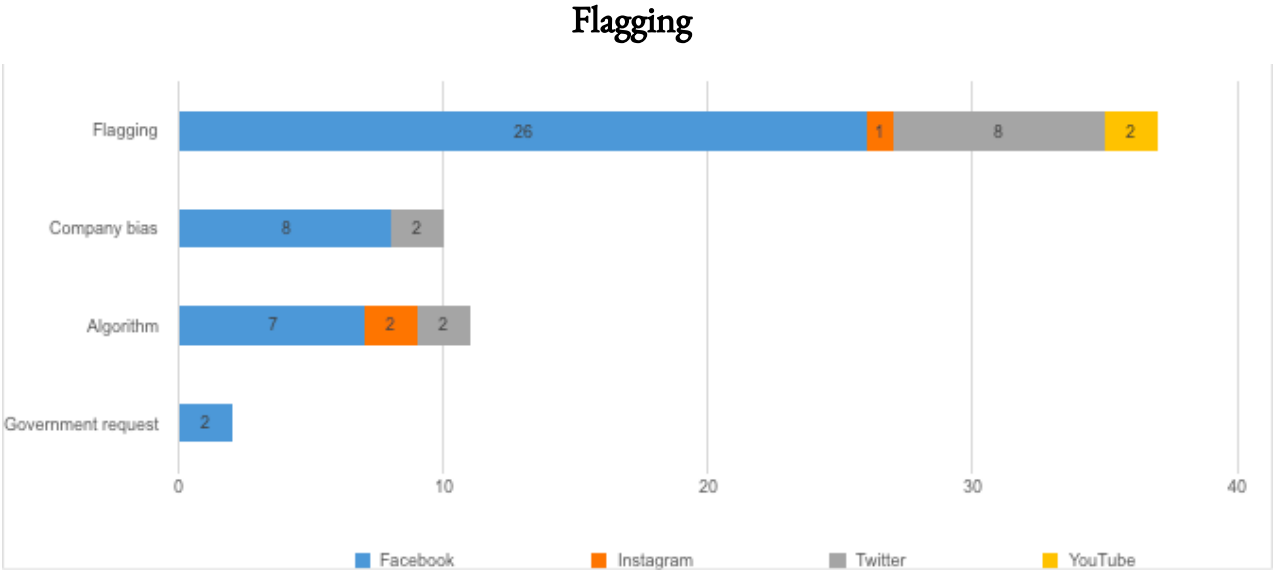
"In only four cases were users able to have their account or content successfully restored, while in almost fifty they reported not even receiving a response."



APPEALS OUTCOMES

Appeals appear to be a particular point of frustration for many users. For example, one respondent said of their experience: "It's ridiculous. I don't feel that I had any appeal process. I submitted two messages/appeals, and did not receive a response to either of them." Another person said: "It is opaque. Completely a cover for saying 'no' with no actual human involvement." A third called the appeals process "futile and pointless."

Ultimately, this reflects a critical need for greater clarity from the companies on what options are available to users when their content and accounts are taken down, and more accountability to help them get back online. User reports reflect a deep distrust in the companies to act in the user's best interest, and a frustration with what they perceive to be insufficient resources and attention paid to this issue.



WHY USERS THOUGHT CONTENT WAS REMOVED

Many users identified flagging as the primary source through which their content is taken down. Often, they reported that they had been targeted by other users, sometimes multiple times, using the reporting features that platforms provide as a mode of user protection in a harmful manner. On occasion, a post was flagged for one reason (such as nudity) and taken down for another (such as violent/graphic content). In addition, there is evidence that governments are making requests for content removals that are taken down due to Terms of Service violations. Twitter recently began submitting reports on these takedowns to Lumen.[13]

Though this was not a question we asked directly, user responses suggest that flagging was more frequently identified as the source of the content takedown. Users also believed that algorithms were responsible for takedowns, as well as political bias exerted by companies over the content that appears on their platforms.

The identification of flagging as a mode of censorship by users reflects the arguments of researchers Kate Crawford and Tarleton Gillespie, who note that while flags have become a ubiquitous mechanism of

[13] "Providing More #transparency into Legal Requests to Remove Content | Twitter Blogs." Providing More #transparency into Legal Requests to Remove Content | Twitter Blogs. February 19, 2016. Accessed March 29, 2016. https://blog.twitter.com/2016/providing-more-transparency-into-legal-requests-to-remove-content.

governance, they are not a direct or uncomplicated representation of community sentiment.[14] Responding to flagged content is also a heavy task reliant on the readiness of human labor to perform at high volume. This leaves the practice of flagging to abuse in which they serve as a mode of targeting individuals and shutting off debate as much as they serve as a tool for combating online harassment.

While flags do not unproblematically lead directly to shutting down speech—in general they do go through additional levels of filtering—the cases illustrate that there are many instances where complex issues slip through the cracks. As Crawford & Gillespie note, the consequences of flagging impact not only the individual who posts and the individual who views, but the broader public life constituted by their interaction.[15]

5. CONCLUSIONS AND RECOMMENDATIONS

The submissions received so far by Onlinecensorship.org have been illuminating. Although our initial sample size is relatively small, those that we have received have allowed us to identify a number of places where users are encountering difficulties and prepare a set of basic recommendations for companies on how to alleviate those concerns.

- Often the communities that are most impacted by online censorship are also the most marginalized—so the people that are censored are also those that are least likely to be heard. Each of the six social media platforms contained in our report are headquartered in the United States, but have cultivated a global reach. While some companies have made strides in opening appropriate regional offices or working with local NGOs, others could greatly improve in this area.
    - **We recommend that companies work with local communities to ensure that their global policies reflect linguistic diversity and other local needs.**
- Several users reported having content taken down that seems to adhere to companies' community guidelines. As more users join social media platforms, the volume of content posted by them will only increase, making effective content moderation a growing challenge for the future.
    - **We recommend that companies devote greater human resources to their internal content moderation teams, continue to broaden their geographic and language reach, and pay particular attention to increasing quality control.**
- A majority of users reported that appeals processes were difficult or nonexistent, and that when they did appeal, their concerns went unheard.

[14] Kate Crawford and Tarleton Gillespie, "What is a flag for? Social media tools and the vocabulary of reporting," *New Media & Society* (2016) vol. 18, 3: 413. Accessed March 23, 2016. doi: 10.1177/1461444814543163.
[15] ibid

    - **We implore companies to improve their appeals processes to ensure that users whose content has been erroneously taken down can easily restore it.**
- Users are typically unable to tell when their content has been removed as a result of a government request. Twitter submits every legal takedown request to Lumen, allowing users and the media to understand when content is being taken down at the behest of government entities, rather than user flagging.
    - **We recommend that other companies follow Twitter's lead and submit government takedown requests to Lumen.**
- Users who violate certain standards report that they are subject to bans of up to thirty days on some platforms. In some cases, these platform bans result in users being cut off from external platforms (third party platforms like Tinder or Spotify use Facebook to authenticate user access) as well as being prohibited from administering certain other content (such as Facebook Pages, which may be used for professional purposes).
    - **We recommend that companies review how users are affected by such bans and make changes to ensure that their professional lives and access to third-party platforms are not impacted.**
- Some policies—such as Facebook's "real name" requirement and the nudity policies imposed by several companies—are perceived by users to be discriminatory, particularly when enforced in private groups or friends-only posts.
    - **We recommend that companies review these policies and consider changing how they are implemented, particularly in non-public spaces (e.g., "secret' Facebook groups or private Instagram accounts).**

FURTHER READING

- Ammori, Marvin. "The 'New' New York Times: Free Speech Lawyering in the Age of Google and Twitter." 127 *Harvard Law Review* 2259 (2014). http://harvardlawreview.org/2014/06/the-new-new-york-times-free-speech-lawyering-in-the-age-of-google-and-twitter/.
- Dewey, Caitlin. "The big myth Facebook needs everyone to believe." *Washington Post*. January 28, 2016. https://www.washingtonpost.com/news/the-intersect/wp/2016/01/28/the-big-myth-facebook-needs-everyone-to-believe/.
- Heins, Marjorie. "The Brave New World of Social Media Censorship." 127 *Harvard Law Review* F.325 (2014). http://harvardlawreview.org/2014/06/the-brave-new-world-of-social-media-censorship/.
- Onlinecensorship.org. "How to Appeal." https://onlinecensorship.org/resources/how-to-appeal
- Ranking Digital Rights. "Corporate Index." https://rankingdigitalrights.org/index2015/
- Taylor, Emily. "The Privatization of Human Rights: Illusions of Consent, Automation and Neutrality." *Global Commission on Internet Governance* (2016). https://ourinternet.org/publication/the-privatization-of-human-rights-illusions-of-consent-automation-and-neutrality/.
- York, Jillian C. "Policing Content in the Quasi-Public Sphere." *OpenNet Initiative* (2010). https://opennet.net/policing-content-quasi-public-sphere.
- York, Jillian C. "Corporate Censorship Is Untouched by the First Amendment." The New York *Times*. August 19, 2014. http://www.nytimes.com/roomfordebate/2013/08/19/can-free-speech-and-internet-filters-co-exist/corporate-censorsip-is-untouched-by-the-first-amendment.