online CENSORSHIP.org

CENSORSHIP

IN CONTEXT

Insights from Crowdsourced Data on Social Media Censorship

By

JESSICA **ANDERSON**
KIM **CARLSON**
MATTHEW **STENDER**
SARAH MYERS **WEST**
JILLIAN C. **YORK**

ELECTRONIC FRONTIER FOUNDATION

VISUALIZING**IMPACT**

This report covers eight months of user-sourced data gathered between March 2016 and October 2016 about 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.

This release follows our inaugural report, "*Unfriending Censorship: Insights from four months of crowdsourced data on social media censorship*," published in March 2016.

Onlinecensorship.org is a collaboration between the Electronic Frontier Foundation (EFF) and Visualizing Impact (VI). EFF is an international non-profit organization with nearly 30,000 members worldwide from over 100 countries, dedicated to the protection of everyone's privacy, freedom of expression, and association. 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

jillian@eff.org | **ramzi@visualizingimpact.org**

Editorial Assistance from **Corey H. Abramson** and **Danny O'Brien**
Design by **Beatrice Jin** and **David Schnitman**

TABLE OF CONTENTS

A YEAR IN ONLINE CENSORSHIP
2016 REVIEW

"THE WORLD'S MOST POWERFUL EDITOR"

Editor-in-chief of *Aftenposten*, Espen Egil Hansen, describing Mark Zuckerberg

142

Media articles on content moderation collected by the Onlinecensorship.org team in 2016

NUDITY
REAL NAME
INAPPROPRIATE CONTENT

Most commonly cited policy violations by Onlincensorship.org users

230

Reports of takedowns received by Onlinecensorship.org between March and October 2016

ARABIC
ENGLISH
SPANISH

Languages in which you can access Onlinecensorship.org

30 DAYS

Maximum period of time a user can be automatically banned on Facebook

32

Reports Onlinecensorship.org received relating to the US elections

103

Number of Onlinecensorship.org users that reported appealing to restore their content

"OUR TEAM PROCESSES MILLIONS OF REPORTS EACH WEEK"

Consistent refrain in Facebook apologies for content moderation issues

"KAFKAESQUE"

Onlinecensorship.org user's description of the experience of being censored on social media

18%

of reports to Onlinecensorship.org dealt with non-English content

6

Number of recommendations Onlinecensorship.org published for companies

TERROR OF WAR

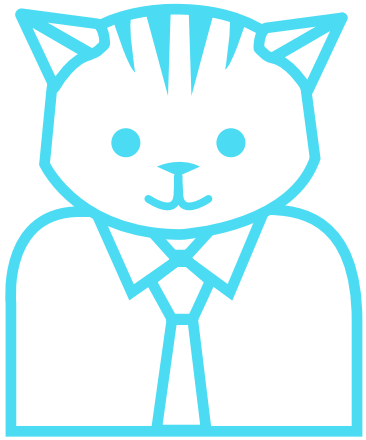
Most high-profile censorship subject of the year

4.75B

Estimated number of Facebook posts per day

3

Number of company policy shifts touched upon in November Onlinecensorship.org report



A CAT IN A BUSINESS SUIT

Our pick for the most bizarre censorship subject of the year

KEY MOMENTS 2016

TIMELINE OF INCIDENTS AND POLICY SHIFTS

January

FACEBOOK CENSORS A FAMED STATUE OF A MERMAID

February

FACEBOOK APOLOGIZES FOR REMOVING A POST BY PROMINENT FILIPINO JOURNALIST ED LINGAO

TWITTER ANNOUNCES IT WILL INCLUDE TERMS OF SERVICE REMOVALS MADE BY LEGAL REQUEST IN ITS TRANSPARENCY REPORTS

March

ABORIGINAL FEMINIST CELESTE LIDDLE RECEIVES A THIRD BAN FROM FACEBOOK FOR POSTS CONTAINING NUDITY

ONLINECENSORSHIP.ORG LAUNCHES ITS FIRST REPORT

April

FACEBOOK LIVE IS RELEASED

May

FACEBOOK SHIFTS TO USING ALGORITHMICALLY CURATED TRENDING TOPICS AFTER ACCUSATIONS OF BIAS IN ITS TREND CURATION

TWITTER SUSPENDS RAPPER AND SINGER AZEALIA BANKS AFTER SHE DIRECTED A SERIES OF RACIST TWEETS AT FORMER ONE DIRECTION SINGER ZAYN MALIK

FACEBOOK, TWITTER, AND YOUTUBE ANNOUNCE A CODE OF CONDUCT AGREEMENT WITH THE EUROPEAN COMMISSION

June

INSTAGRAM REMOVES A PHOTO OF A PLUS SIZE MODEL

July

UNDER PRESSURE TO STOP HARRASSMENT, TWITTER PERMANENTLY BANS MILO YIANNOPOULOS FOR ATTACKING LESLIE JONES

LIVE VIDEO OF PHILANDO CASTILE SHOOTING DISAPPEARS FROM FACEBOOK

August

FACEBOOK REMOVES MEME CALLING CONVICTED RAPIST BROCK TURNER A RAPIST

TWITTER SUSPENDS 235,000 ACCOUNTS OVER EXTREMISM

September

FACEBOOK DEMANDS *TERROR OF WAR* PHOTO BE REMOVED OR PIXELATED

FACEBOOK APOLOGIZES FOR REMOVING PAGES OF SEVERAL PROMINENT PALESTINIAN JOURNALISM OUTLETS IN THE WAKE OF HIGH LEVEL MEETINGS BETWEEN FACEBOOK AND ISRAEL

TWITTER INTRODUCES NEW APPS FOR LIVE STREAMING

October

A PHOTO OF A CAT IN A BUSINESS SUIT ALLEGEDLY LEADS TO A FACEBOOK ACCOUNT SUSPENSION

FACEBOOK ANNOUNCES IT WILL SEEK INPUT FROM THE COMMUNITY AND PARTNERS ON MAKING EXCEPTIONS TO ITS TERMS OF SERVICE FOR MATERIAL "IN THE PUBLIC INTEREST"

November

IN THE WAKE OF THE US ELECTION, MARK ZUCKERBERG PROMISES TO PROPOSE SOLUTIONS FOR TACKLING FAKE NEWS AND HOAXES SHARED ON FACEBOOK.

INTRODUCTION

Over the past year, the issue of content moderation has drawn increased attention, resulting in more visible public debate over the role companies play in regulating speech.

A year after its launch, Onlinecensorship.org seeks to contribute to this debate by asking users who have had their content censored by social media platforms to describe their experiences. In this, our second report, we focus on providing insight from our user reports to develop a concrete accountability framework for social media content moderation.

CONTEXT IS KEY

Since our inaugural report[1] in March of this year, media coverage of online censorship seems to have reached an all-time high. The Onlinecensorship.org team is tracking media coverage in order to better understand the focus and direction of this debate. Since 2015, we've collected 208 articles covering censorship on social media. This includes 66 articles from 2015, 142 so far in 2016, and 99 published since our first report in March 2016.

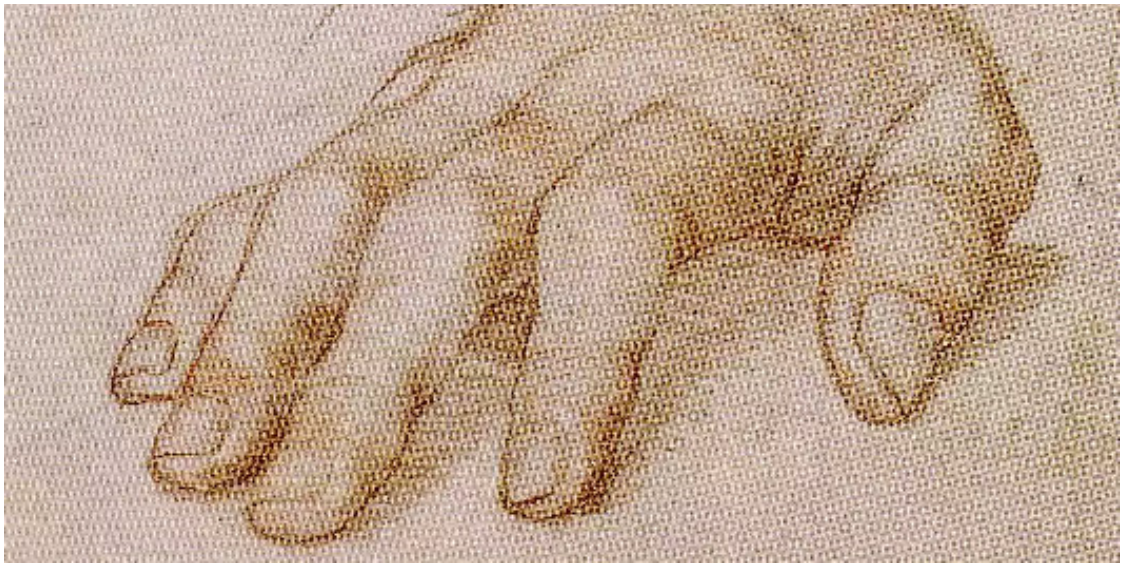
From these articles, we extracted 294 examples of censorious content moderation.[2] We coded the 294 censorship examples by platform, reason for censorship, and type of content affected, and applied a tagging system to parse out frequently discussed issues and communities. Some of the insights from our media analysis include:

- Facebook was the most covered platform by a wide margin. 74% of censorship examples reported in the articles we reviewed concerned Facebook, followed by Instagram (16%) and Twitter (7%).

- Removal of photos, making up 31% of examples, generated the most coverage. Other common subjects of discussion were account suspensions (28%) and post removals (12%).

- 47% of examples concerned the moderation of content for nudity or sexually explicit content or information. Other examples referenced applications of policies on hate speech (10%), graphic content (8%), and inauthentic identity (5%). The reason for censorship was unclear in 23% of examples.

- Tags associated with nudity, in order of frequency, were: female nudity, art, women's bodies (non-nude), indigenous people, breastfeeding, LGBTQI, Kim Phúc, menstruation, and health.



Study of Right Hand of Erasmus of Rotterdam, Hans Holbein the Younger. This 16th century image was removed from Facebook in August for violating its community guidelines. The decision was reversed after protests from art fans.

[1] Onlinecensorship.org, "Unfriending Censorship: Insights from four months of crowdsourced data on social media censorship," March 31, 2016, https://onlinecensorship.org/news-and-analysis/onlinecensorship-org-launches-first-report-download.

[2] The 294 examples contain duplications of examples that were referenced in more than one article. We did not eliminate these duplications because they helped weight our dataset toward the issues that have captured the most media attention.

- Communities discussed included: women, artists, LGBTQI, Indigenous people, Palestinians, African Americans (in the context of police brutality), Muslims, American conservatives, journalists, breastfeeding mothers, Kashmiris, Kurds, and refugees.

- A common theme throughout the media coverage was the challenge social media companies face in employing nuance when enforcing their content policies.

In this report, we add value to existing media coverage of content moderation from two angles. First, we analyze how several high profile cases have catalyzed changes in both public perception and company policies. Second, we highlight emerging issues not yet prominent in the public discussion.

From Newsworthy Content to Fake News

2016's most high-profile censorship incident occurred when Norwegian author Tom Egeland posted a series of photos documenting the history of military conflicts. Contained within the photo series was Nick Ut's famous Vietnam War-era photograph; entitled The Terror of War, the photograph depicts nine-year-old Kim Phúc fleeing naked from a napalm attack.

Facebook removed the image,[3] prompting Espen Egil Hansen—the editor-in-chief of Norwegian publication *Aftenposten*—to post a front-page open letter to Facebook CEO Mark Zuckerberg in which he called Zuckerberg "the world's most powerful editor." Hansen called on him to revise the company's policies to be more inclusive of newsworthy content that might otherwise violate Facebook's "Community Standards."[4]

Facebook reinstated the photo, and in a statement to the *Guardian*,[5] stated:

> **"After hearing from our community, we looked again at how our Community Standards were applied in this case. An image of a naked child would normally be presumed to violate our Community Standards, and in some countries might even qualify as child pornography. In this case, we recognize the history and global importance of this image in documenting a particular moment in time."**

Less than two months after the takedown and its ensuing media storm, Facebook announced a new change in policy,[6] stating in a press release that the company would "begin allowing more items that people find newsworthy, significant, or important to the public interest—even if they might otherwise violate our standards." Facebook also announced an intent to work with its community and partners to "explore exactly how to do this, both through new tools and approaches to enforcement" and stated their aim to "allow more images and stories without posing safety risks or showing graphic images to minors and others who do not want to see them."

---

[3] Kevin Lui, "Mark Zuckerberg Has Been Blasted Over Facebook's Censorship of the 'Napalm Girl' Photo," *TIME*, 9 September 2016, http://time.com/4484905/facebook-censorship-napalm-girl-newspaper-editor-norway/.

[4] Espen Egil Hansen, "Dear Mark. I am writing this to inform you that I shall not comply with your requirement to remove this picture," *Aftenposten*, 9 September, 2016, http://www.aftenposten.no/meninger/kommentar/Dear-Mark-I-am-writing-this-to-inform-you-that-I-shall-not-comply-with-your-requirement-to-remove-this-picture-604156b.html

[5] Sam Levin et al., "Facebook backs down from 'napalm girl' censorship and reinstates photo," *the Guardian*, 9 September 2016, https://www.theguardian.com/technology/2016/sep/09/facebook-reinstates-napalm-girl-photo.

[6] Facebook, "Input from Community and Partners on our Community Standards," 21 October 2016, http://newsroom.fb.com/news/2016/10/input-from-community-and-partners-on-our-community-standards/.

The announcement is a step forward for the company, and brings into focus the difficult balancing act faced by social media companies: allowing the free flow of information while also creating and implementing policies that satisfy their varied (and often at odds) constituencies. Even when such policies are well-crafted, companies face even greater challenges moderating content at scale.

Issues of context in content moderation decisions have raised important questions: what editorial responsibilities do social media companies have? CEO Mark Zuckerberg plainly stated several months ago that Facebook is "a tech company, not a media company," further noting that "we build the tools, we do not produce any content." [7] Despite his insistence, a recent *Reuters* investigation found that "an elite group of at least five senior executives regularly directs content policy and makes editorial judgment calls, particularly in high-profile controversies…" [8]

The sharing of fake news and hoaxes on Facebook in the lead up to the election of Donald Trump brought the company's influence under scrutiny again. Responding to critiques that it may have influenced the election, the company has said it is "just one of many ways people received their information—and was one of the many ways people connected with their leaders, engaged in the political process and shared their views." [9] Our reports suggest that this statement does not adequately capture the degree to which some users put Facebook at the center of their communications ecosystem. Though we agree with Facebook CEO Mark Zuckerberg's comment that the company "must be extremely cautious about becoming arbiters of truth,"[10] we argue that embracing, rather than denying, this role would lead to greater responsibility and accountability by the company.

Critics of the company point to Facebook's direct partnerships with publishers and advertisers and the power of their algorithmic control over what is visible to their users to make the case that Facebook is an active editor of content, and has a responsibility to ensure its editorial decisions are in the public interest.[11] As the reports we received demonstrate, many Facebook users concur with this assessment of social media companies' responsibilities. Their comments point toward specific policy and design changes that may contribute to improving companies' accountability to the public.

Hate Speech

Another frequently cited issue in press coverage of content moderation over the past year has been the challenge of regulating hate speech. This was particularly the case in July, when Leslie Jones was attacked on Twitter with racist and sexist content that violated the company's Twitter "Rules".[12] Twitter responded to the attacks by banning Milo Yiannopoulos permanently from the platform for his efforts to instigate the attacks on Jones. Yiannopoulos' ban was only one of a few instances in which a user has been permanently banned from Twitter. [13]

[7] Giulia Segreti, "Facebook CEO says group will not become a media company," *Reuters*, 29 August 2016, http://www.reuters.com/article/us-facebook-zuckerberg-idUSKCN1141WN.
[8] Kristina Cooke et al., "Facebook executives feel the heat of content controversies," *Reuters*, 28 October 2016, http://mobile.reuters.com/article/idUSKCN12S0D3.
[9] Mike Isaac, "Facebook, in Cross Hairs After Election, Is Said to Question Its Influence," New York Times, 12 November 2016, http://www.nytimes.com/2016/11/14/technology/facebook-is-said-to-question-its-influence-in-election.html.
[10] Ibid.
[11] Mathew Ingram, "Sorry Mark Zuckerberg, But Facebook Is Definitely a Media Company," *Fortune*, 30 August 2016, http://fortune.com/2016/08/30/facebook-media-company/.
[12] Kristen V. Brown, "How a racist, sexist hate mob forced Leslie Jones off Twitter," *Fusion*, 19 July 2016, http://fusion.net/story/327103/leslie-jones-twitter-racism/.
[13] Elle Hunt, "Milo Yiannopoulos, rightwing writer, permanently banned from Twitter," *the Guardian*, 20 July 2016, https://www.theguardian.com/technology/2016/jul/20/milo-yiannopoulos-nero-permanently-banned-twitter.

While all of the companies we researched are based in the United States—where there are few direct prohibitions on hate speech—a majority of countries regulate hateful or discriminatory speech in some manner. How "hate speech" is defined may differ, however, from one country to the next, making regulation of it on global platforms difficult. Most companies go beyond legal requirements in their efforts to create a safe and welcoming environment for users.

Of the companies we've studied, all six have policies that pertain to hate speech. Twitter's "Rules",[14] for example, state that users "may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease." The rules further state that accounts whose primary purpose is to incite harm on the basis of these categories are not allowed.

Similarly, Facebook's "Community Standards" state that content that directly attacks people based on the same list of categories (minus "age") as Twitter is removed, and that organizations and people who promote hatred against these protected groups are not allowed a presence on the platform. Google's rules are comparable, but expressly note that "there is a fine line between what is and what is not considered to be hate speech. For instance, it is generally okay to criticize a nation-state, but not okay to post malicious hateful comments about a group of people solely based on their ethnicity."

Twitter has for some time faced criticism from users who feel that the company doesn't go far enough in monitoring and moderating hateful speech and harassment. In 2014, after months of criticism, Twitter partnered with Women Action Media (WAM!) to identify gendered harassment and devise solutions for dealing with it.[15] During the 2016 United States election season, numerous journalists reported being targeted by Donald Trump supporters on the platform, often with anti-Semitic imagery.[16]
Nevertheless, the company has taken steps against harassment, notably against several high-profile users. In May 2016, rapper and singer Azealia Banks was suspended from the platform after she directed a series of racist tweets at former One Direction singer Zayn Malik.[17] Breitbart writer Milo Yiannopoulos was first "punished" by the removal of his verified status,[18] then later suspended after harassing actress Leslie Jones.[19]

Despite these high-profile suspensions, recent media reports suggest that harassment on the platform dissuaded two potential bidders from their interest in buying the company.[20] [21]

[14] Twitter, "The Twitter Rules," https://support.twitter.com/articles/18311com/2016/08/30/facebook-media-company/.

[15] Amanda Marcotte, "Can These Feminists Fix Twitter's Harassment Problem?," *Slate*, 7 November 2014, http://www.slate.com/blogs/xx_factor/2014/11/07/women_action_media_and_twitter_team_up_to_fight_sexist_harassment_online.html.

[16] *NPR*, "Harassed On Twitter: 'People Need To Know The Reality Of What It's Like Out There'," 26 October 2016, http://www.npr.org/2016/10/26/499440089/harassed-on-twitter-people-need-to-know-the-reality-of-what-its-like-out-there.

[17] Nicky Woolf, "Azealia Banks suspended from Twitter following racist attack on Zayn Malik," *the Guardian*, 13 May 2016, https://www.theguardian.com/music/2016/may/12/azealia-banks-suspended-twitter-racist-rant-zayn-malik.

[18] Alex Kantrowitz, "Twitter Unverifies Writer Amid Speech Wars," *Buzzfeed*, 9 January 2016, https://www.buzzfeed.com/alexkantrowitz/twitter-unverifies-milo-yiannopoulos-leading-to-speech-polic?utm_term=.kewPrk9MX#.tbbLN7rmJ.

[19] Marta Cooper and Joon Ian Wong, "Milo Yiannopoulos has been annoying people for a very, very long time," *Quartz*, 20 July 2016, http://qz.com/737148/milo-yiannopoulos-has-been-annoying-people-for-a-very-very-long-time/.

[20] *Fortune*, "Here's why Disney and Salesforce.com dropped their bids for Twitter," 18 October 2016, http://fortune.com/2016/10/18/twitter-disney-salesforce/

[21] Rich McCormick, "Twitter's reputation for abuse is putting off potential suitors," *the Verge*, 19 October 2016, http://www.theverge.com/2016/10/19/13328380/twitter-abuse-disney-salesforce-bids-pulled.

There have been several high profile instances of images of police brutality being removed by social media platforms.

Further, complaints remain that Twitter often fails to ban odious harassers. The swiftness with which they responded to high-profile complaints about Banks and Yiannopoulos seems unrepresentative of their broader response to harassment and hate speech: for example, a report by WAM! on harassment on Twitter found that the vast majority, 79.7%, of those reporting cases experienced harassment that had been going on for a week or longer.[22]

New Tools, New Censorship

New technology in social media has created new expectations for content moderation and made the inclusion of context in those decisions more urgent than ever. Social movements around the world—from Black Lives Matter in the United States to the fight for Kurdish independence in the Middle East—have long relied on social media to organize and spread their messages. New livestreaming tools such as Facebook Live and Twitter's Periscope have made it easier for these movements to document what's happening around them, including violence and atrocities, but have also presented companies with the new challenges of moderating content in real-time, often at the behest of governments.

The emergence of Facebook Live in April of 2016 was shortly followed by its first widely-reported case of censorship, when the video of police shooting of Philando Castile in Minnesota using the tool by Castile's girlfriend, Diamond Reynolds, was taken down.[23] Although Facebook apologized and later restored the video, in August a second incident occurred in which the Baltimore County police department successfully ordered Facebook to deactivate the account of Korryn Gaines while she was involved in a standoff with police, who later shot and killed her.[24]

The introduction of live streaming video on Twitter and Facebook means that these platforms are now being used to broadcast newsworthy content in real time. This also means that decisions made to take live content down may have a greater impact on the public visibility of the broadcast: losing viewers due to an error in the interpretation of a policy may mean that the story itself may fail to gain traction.

---

[22] J. Nathan Mathias et al., "Reporting, Reviewing, and Responding to Harassment on Twitter." Women, Action, and the Media, 13 May 2015, http://womenactionmedia.org/twitter-report.

[23] Lucy Tiven, "Mark Zuckerberg Takes a Stance on the Video of Philando Castile's Death," *attn:*, 10 July 2016, http://www.attn.com/stories/9809/mark-zuckerberg-addressed-video-philando-castiles-death.

[24] Baynard Woods, "Facebook deactivated Korryn Gaines' account during standoff, police say," *the Guardian*, 3 August 2016, https://www.theguardian.com/us-news/2016/aug/03/korryn-gaines-facebook-account-baltimore-police.

Privatized Enforcement

A final theme that emerged from the media coverage is the issue of privatized enforcement. In the United States and elsewhere, governments are increasingly relying on companies to act as judge, jury, and executors of the law.

In December 2015, an agreement was reached between the German government and Twitter, Facebook, and Google that requires the companies to remove illegal hate speech within twenty-four hours. Germany's Justice Minister has stated that "[t]he new agreement makes it easier for users and anti-racism groups to report hate speech to specialist teams at the three companies."[25]

In May 2016, the European Commission announced a "code of conduct" with four "IT companies" (Facebook, Twitter, YouTube, and Microsoft) that includes several public commitments "to combat the spread of illegal hate speech online in Europe."[26] Similar to the agreement with Germany, the code of conduct focuses on speech that is illegal in Europe, requiring companies to review flags of such content and remove it within 24 hours "and remove or disable access to such content, if necessary." The code of conduct also asks the IT companies to: "intensify cooperation" between themselves; "educate and raise awareness with their users" about types of content not permitted"; and cooperate with civil society organizations to "deliver best practice training on countering hateful rhetoric and prejudice."[27]

The European code of conduct has been criticized by civil society organizations for a number of reasons, including its lack of a multi-stakeholder approach. The European Digital Rights Initiative (EDRi) claimed that civil society organizations were "systematically excluded" from the negotiations that led to the development of the code of conduct and along with several other organizations, blocked from future discussions held under the banner of the "EU Internet Forum," the body set up to counter "terrorist activity" and "hate speech" online.[28]

EDRi further criticized the European Commission's approach to privatized enforcement, writing: [29]

> **"Sadly, the discussions keep running in circles: a public policy problem is identified; public authorities put pressure on companies to solve it (even after the adoption of the Code, countries like Germany keep putting pressure on social media companies to do the job of a public authority); the content is the target, not the author(s) of the content; Member states do not investigate or prosecute; the Commission does not take responsibility; the root of the problem is not solved. Only if the problem is the need to be seen to be doing "something", the problem is solved."**

In addition to the European agreements, news reports of an agreement between Israel and Facebook have persisted throughout the year.[30] While Facebook has not confirmed reports of an agreement, a senior Israeli cabinet minister announced in September that the government had "agreed to work together to determine how to tackle incite-

[25] Reuters, "Facebook, Google, Twitter agree to delete hate speech in 24 hours: Germany," 15 December 2015, http://www.reuters.com/article/us-germany-internet-idUSKBN0TY27R20151215.

[26] European Commission press release, "European Commission and IT Companies announce Code of Conduct on illegal online hate speech," 31 May 2016, http://europa.eu/rapid/press-release_IP-16-1937_en.htm.

[27] Ibid.

[28] European Digital Rights Initiative, "EDRi and Access Now withdraw from the EU Commission IT Forum discussions," 31 May 2016, https://edri.org/edri-access-now-withdraw-eu-commission-forum-discussions/.

[29] Ibid.

[30] Mazal Mualem, "Can Israel twist Facebook's arm?," Al Monitor, 26 February 2016, http://www.al-monitor.com/pulse/originals/2016/02/israel-minister-erdan-facebook-war-incitement-palestine.html.

ment on the social media network," according to a report from the *Guardian*.[31] Later that month, Facebook disabled the accounts of several editors and executives at two prominent Palestinian publications,[32] a move for which the company later apologized.[33] In response, a number of civil society organizations issued a statement in opposition to the agreement.[34]

Agreements between social media platforms and governments to remove content fall into the growing class of "shadow regulations" on the Internet, which bypass the transparency and due process expected in public policy,[35] and run afoul of principles laid out by civil society organizations for corporate policies, and have been rightly criticized as such. A recent letter from seventy-three rights organizations called on Facebook to clarify its policy on removing content, "especially human rights documentation," at the request of government actors, noting the "increasingly central role in controlling media that circulates through the public sphere."[36]

The letter further recommends a public appeals platform where users can appeal content decisions.

The Manila Principles,[37] released in early 2015, lay out a set of six basic guidelines for intermediaries, including social media companies. The first two principles state that "[i]ntermediaries should be shielded from liability for third party content" and that "[c]ontent must not be required to be restricted without an order by a judicial authority," while the other guidelines focus on due process, necessity and proportionality of content restriction orders, and transparency and accountability. A background paper released by the coalition supporting the Principles further defines the roles and responsibilities of companies.[38]

The Principles follow numerous efforts over the years to develop a framework for understanding the role that companies play in the greater regulatory ecosystem. One of the first major efforts to this end was the creation of the UN Guiding Principles on Business and Human Rights. Endorsed by the UN Human Rights Council in 2011, these Principles lay out a basic framework for companies to understand their role and responsibilities. [39]

Rebecca MacKinnon's 2012 *Consent of the Networked* also called for a "Magna Carta for the Internet" and emphasized the idea of user consent. MacKinnon later developed Ranking Digital Rights,[40] a project which ranks companies on their disclosed commitments, policies and practices that affect user privacy and freedom of expression. The methodology for the rankings take

[31] Associated Press, "Facebook and Israel to work to monitor posts that incite violence," 12 September 2016, https://www.theguardian.com/technology/2016/sep/12/facebook-israel-monitor-posts-incite-violence-social-media?CMP=twt_gu

[32] Amar Toor, "Facebook accused of censoring Palestinian journalists," *the Verge*, 26 September 2016, http://www.theverge.com/2016/9/26/13055862/facebook-israel-palestinian-journalists-censorship.

[33] Ali Abunimah, "Facebook apologizes for disabling Palestinian journalists' accounts," *Electronic Intifada*, 24 September 2016, https://electronicintifada.net/blogs/ali-abunimah/facebook-apologizes-disabling-palestinian-journalists-accounts

[34] SMEX, "Civil Society Statement: Voluntary Agreements between Israel and Social Media Companies Threaten Free Expression," 3 October 2016, http://www.smex.org/civil-society-statement-voluntary-agreements-between-israel-and-social-media-companies-threaten-free-expression/.

[35] Jeremy Malcolm and Mitch Stoltz, "Shadow Regulation: The Back-Room Threat to Digital Rights", 29 September 2016, https://www.eff.org/deeplinks/2016/09/shadow-regulation-back-room-threat-digital-rights.

[36] Civil Society Letter to Facebook, 31 October 2016, http://sumofus.org/wp-content/uploads/2016/10/NewFacebookCensorshipSign-OnLetter.pdf.

[37] Manila Principles on Intermediary Liability, https://www.manilaprinciples.org/.

[38] "The Manila Principles on Intermediary Liability Background Paper," 31 May 2015, https://www.eff.org/files/2015/07/08/manila_principles_background_paper.pdf.

[39] United Nations Office of the High Commissioner for Human Rights, "Guiding Principles on Business and Human Rights," 2011, http://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf.

[40] Ranking Digital Rights, https://rankingdigitalrights.org/.

into account both the UN Guiding Principles as well as the Global Network Initiative's Principles[41] and Implementation Guidelines.[42]

More recently, David Kaye, the United Nations Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, issued a set of recommendations for states and companies dealing with speech regulation. Recognizing that "[o]nline expression is increasingly mediated through private networks and platforms created, maintained and operated by a diverse range of companies…," Kaye's recommendations include: [43]

- States must not require or otherwise pressure the private sector to take steps that unnecessarily or disproportionately interfere with freedom of expression, whether through laws, policies, or extra legal means.

- Any demands, requests and other measures to take down digital content or access customer information must be based on validly enacted law, subject to external and independent oversight, and demonstrate a necessary and proportionate means of achieving one or more aims under article 19 (3) of the International Covenant on Civil and Political Rights. Particularly in the context of regulating the private sector, State laws and policies must be transparently adopted and implemented.

Kaye's recommendations are echoed by Emily Taylor, who urges states to "be vigilant in monitoring the impact private actors have on human rights, redressing imbalances where necessary."[44] In a paper sympathetic to the challenges companies face, Taylor remarks: "Despite sincere commitments to freedom of expression, and the legal incentives to maintain neutral intermediary status, popular web platform providers have become drawn into making decisions to remove or moderate content."

EDRi recognizes the complexity of this issue, asking in a 2014 paper: [45]

> **"How can it be that the digital world that has generated – and continues to generate – such opportunities for human rights can also be used as a tool for such destruction of human rights? The answer is populist reliance on private companies to regulate an environment that does not always lend itself easily to law and law enforcement. It is essential and urgent to stop the erosion of the rule of law, democracy, freedom of assembly, freedom of communication, privacy and legal certainty. As a society, we need to cherish the democratic potential of digital technologies and – even when this is difficult – prevent the silent digital decay of both online and offline rights that the we appear to have taken for granted."**

Each of these frameworks lays out a set of principles for the role of companies in upholding human rights as they develop policies. In this report, we draw on the accounts of users submitting to Onlinecensorship.org in order to operationalize them in the content moderation context. In subsequent sections, we will cover key themes that emerged from these accounts, as well as offer recommendations to the companies based on user experiences.

[41] Global Network Initiative, "Principles," https://globalnetworkinitiative.org/principles/index.php.

[42] Global Network Initiative, "Implementation Guidelines," http://globalnetworkinitiative.org/implementationguidelines/index.php

[43] United Nations, General Assembly, "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression," A/HRC/32/38 (11 May 2016), available from http://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/Privatesectorinthedigitalage.aspx.

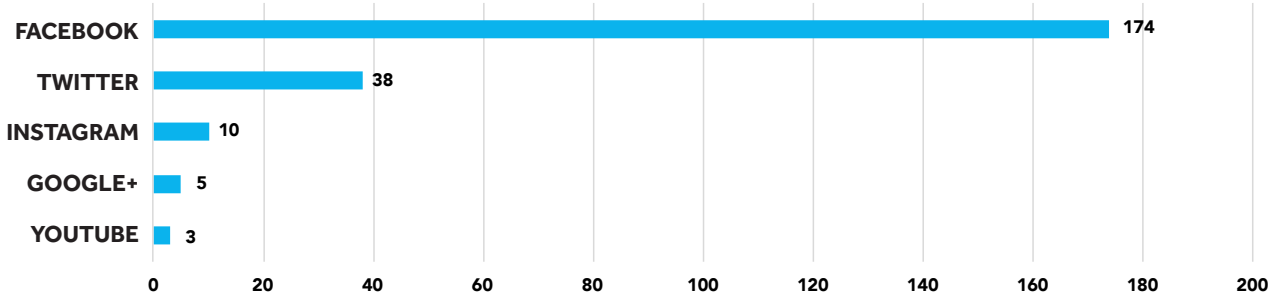
[44] Emily Taylor, "The Privatization of Human Rights: Illusions of Consent, Automation and Neutrality," CIGI/Chatham House, January 2016, available from https://www.ourinternet.org/research/privatization-human-rights-illusions-consent-automation-and-neutrality.

[45] European Digital Rights Initiative, "Human Rights and privatised law enforcement : Abandoning rights - abandoning democracy - abandoning law," 25 February 2014, https://edri.org/wp-content/uploads/2014/02/EDRi_HumanRights_and_PrivLaw_web.pdf.

LEARNING FROM USER EXPERIENCES OF CONTENT MODERATION

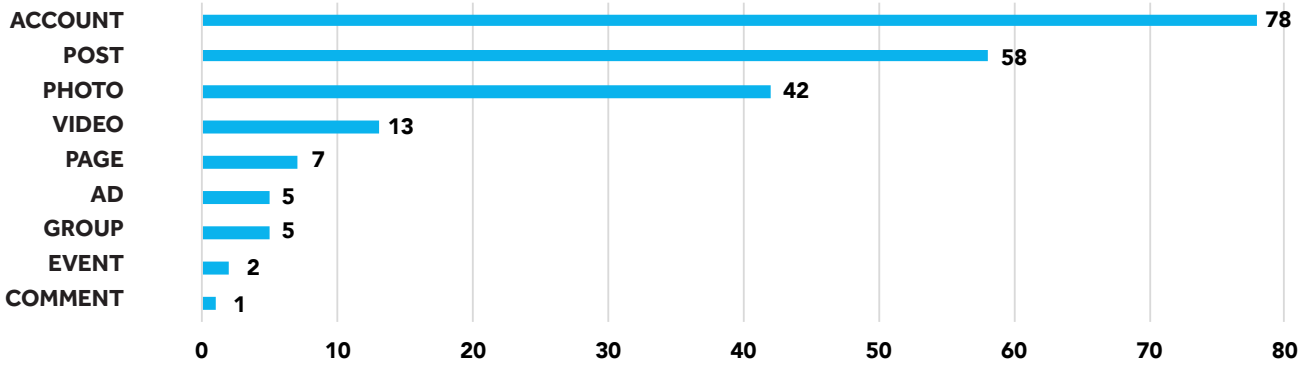
In total, Onlinecensorship.org received 230 reports of takedowns between March and October of 2016. Of these reports, roughly three quarters (75.7%) of the reports pertained to takedowns on Facebook, 16.5% related to content removed from Twitter, 4.4% on Instagram, 2.2% on Google, and 1.3% on YouTube. We did not receive any reports relating to content removals on Flickr during this period.

REPORTS BY PLATFORM



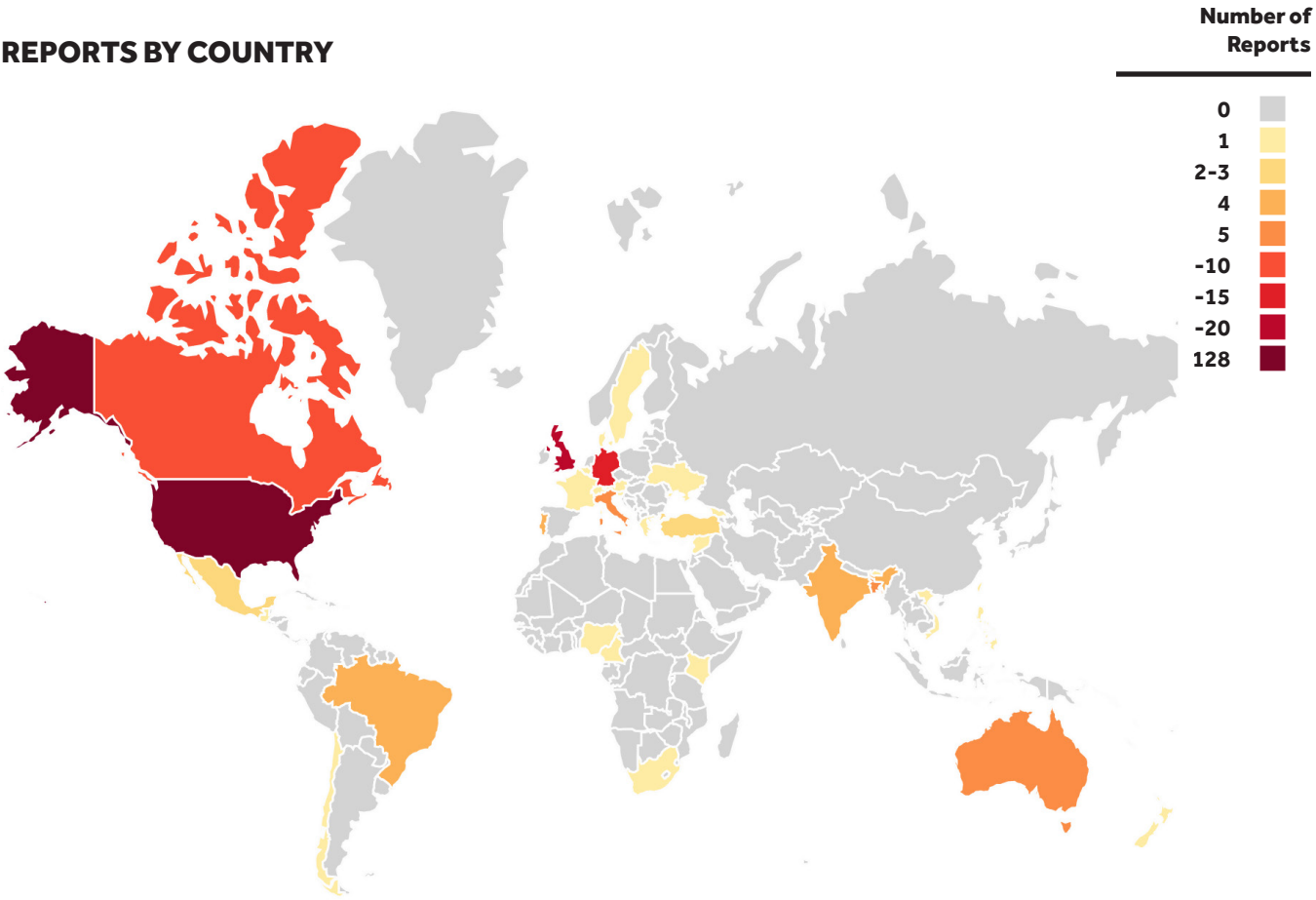
The majority of reports (35.5%) related to account shutdowns. Though we did not collect this information in the report submissions, several Facebook users volunteered that the shutdown of their account was through an automatic ban, which lasted for a fixed period of time ranging from 24 hours to 30 days, and which could not be appealed. 26.4% of the reports related to the takedown of a post, 19.1% to the takedown of a photo, and 5.9% to the removal of a video. 8% of those submitting a report did not provide any detail about what type of content was removed.

REPORTS BY CONTENT TYPE



The vast majority of the reports (82%) related to English language content. Spanish was the second most frequently reported language at 3%, followed by German (2%) and Bengali (2%). Over half (62%) of the reports came from users based in the US, followed by the United Kingdom (9%), Germany (5%) and Canada (5%).

REPORTS BY COUNTRY



Most of the reports either did not indicate a reason for the takedown, or indicated the user was not provided the reason by the company. In many cases, even if they were not given a reason for the takedown, these users did have ideas as to why their content was taken down, which we will discuss in more detail shortly. Of those that did provide a reason for the content takedown, nearly half (47%) related to nudity, all of which occurred on Facebook. 18% related to violations of Facebook's real name policy, and 10% related to "inappropriate content," all of which occurred on Instagram. Twitter and Instagram users were less likely to report receiving a reason for the content takedown than Facebook users, although the higher volume of Facebook reports overall could account for this discrepancy.

US Elections

The US presidential elections figured prominently in the reports we received. We received 32 reports of content takedowns overall relating to the elections, the majority of which were on Twitter and Facebook. Though these takedowns occurred on both sides of the aisle, most of the reports related to content critical of Hillary Clinton, posted by users who either identified as Bernie Sanders supporters or Donald Trump supporters. Several users also reported takedowns of a story about alleged censorship of an interview with Jill Stein. It's important to note that this is not necessarily an indication of a skew in actual number of takedowns—in fact, reporting by the Wall Street Journal suggests that Facebook made adjustments to its evaluation of its community guidelines in order to strive for objectivity in representing offline discourse, leaving up posts by Donald Trump calling for a ban on Muslims that would otherwise qualify as hate speech under the company's com-

munity guidelines.[46] Thus, the accounts represented in this report should be interpreted as a source of insight into the perceptions of the users reporting to Onlinecensorship.org, and not users on the whole.

Many of these users reported extreme frustration with the removal of their content, as they sought to speak their minds and share information about the highly contested election. They were more likely than others to allege political motivations on the part of the companies being the reason for the takedown. Though a minority, a number of these users expressed sentiments along the lines of: "Facebook is engaged in systematic political censorship," suggesting they believed the company was politically biased in favor of Hillary Clinton.

More often, however, they attributed the removal of their content to flagging by other politically motivated users. For example, one user said "I'm guessing their [sic] is either a financial incentive or a conflicted party that wants to scrub twitter of anti-Hillary media."

Generally speaking, the election-related reports we received correlated with a recent Pew survey that found 59% of social media users reported their interactions with those with opposing political views were stressful and frustrating, and that many see political conversations on social media as angrier, less respectful and less civil than in other areas of their life.[47] This may have downstream effects on speech on social media platforms in the form of self-censorship: over half of Pew respondents said they exercised constraint when posting to social media for fear they would be criticized. Many users do so either by outright refraining from voicing controversial opinions, or by moving conversations to gated, more private, inclusive channels like a closed, moderated group or shifting to another platform altogether.[48]

[46] Deepa Seetharaman, "Facebook employees pushed to remove Trump's posts as hate speech," *Wall Street Journal*, 21 October, 2016. http://www.wsj.com/articles/facebook-employees-pushed-to-remove-trump-posts-as-hate-speech-1477075392

[47] Pew Research, "The Political Environment on Social Media," 25 October 2016. http://www.pewinternet.org/2016/10/25/the-political-environment-on-social-media/.

[48] Tasbeeh Herwees, "The Curious Rise of Secret Facebook Groups," *GOOD*, 24 August 2016, https://www.good.is/articles/the-secret-facebook-groups-where-women-commune-and-heal.

Many users expressed frustration about the removal of content they posted about presidential candidates.

"Donald Trump in Reno, Nevada" by Darron Birgenheier is licensed under CC BY-SA 2.0. https://www.flickr.com/photos/darronb/23679921353

Interpretations of Content Takedowns

As indicated in the introduction to this section, many users reported they did not have a clear sense of the reason why their content was taken down–only 60 reports contained a stated or reported reason for the takedown. Twitter users were more likely to suggest they had a vague sense of why their content was taken down than users of other platforms.

Often, the users suggested that the policies outlined by companies were only a proxy for a system that more frequently allows users to shut down one another's speech as a form of retribution. In the words of one Facebook user, "anybody can report you for any reason, and they can block you for ANY reason and when you ask them for specifics, they just send you a standard 'you violated commu[n]ity guidelines'... it's made specifically so people can ELIMINATE speech they find 'offensive' and facebook can pretend like the[y] had nothing to do with it."

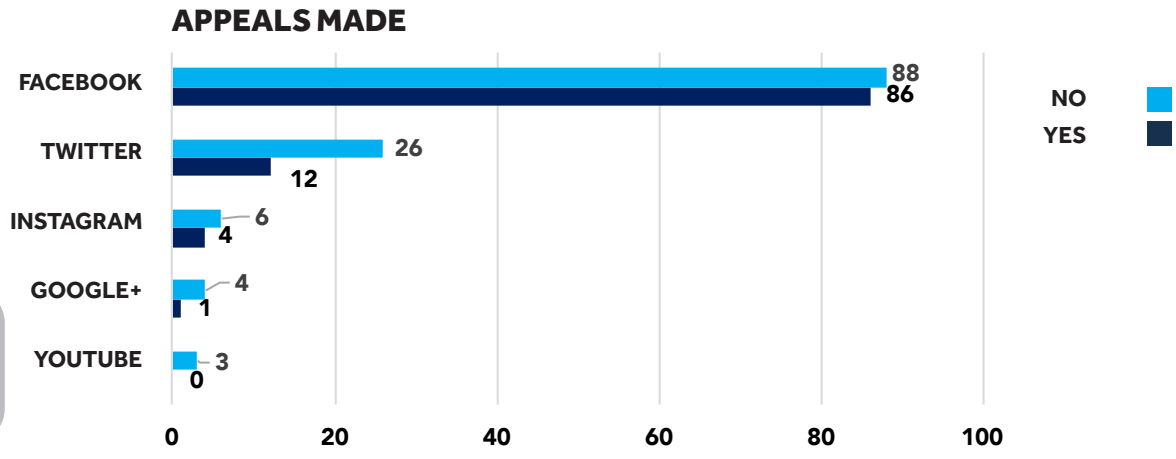
This suggests that users are dissatisfied with the current reporting mechanisms available to them, and distrustful of the level of scrutiny applied to flagged content.

Appeals Processes and User Redress

Many users are opting to appeal the takedown of their content–103 (44.7%) in total during this period. While we were not able to discern the outcomes of most of these appeals, at least 28 of those who did appeal said they did not get any response. Over half of Facebook users submitting reports to Onlinecensorship.org did not appeal.

Some users reported getting contradictory information during the appeals process: one Facebook user said "As you will see from the screen-shot, my account was disabled, I appealed and received a response apologising and saying my account had been reactivated. However, when I tried to log-in [sic], it said they were verifying my ID (but I had not been asked for any ID docs). I replied saying [that] my account had not been reactivated and received a reply saying my account had been disabled. A complete contradiction to the previous reply and now back to square one, except they closed the thread and now I can no longer reply."

Perhaps related to this confusion about how to appeal, a number of users reported turning to other mechanisms in order to try to obtain redress, such as writing to technical support pages and advertiser pages. These users reported they were more likely to get a response from an employee of the company via these mechanisms, but that this did not translate into the successful restoration of their content.



Private Versus Public Content Takedowns

Some users reported feeling it was unfair that they had their content taken down from private groups or sent via private messages, indicating that they draw a distinction between content they post publicly and content they post among consenting users privately.

One user described this frustration at length: "I run a closed and 'secret' group called [name redacted]. Yes we do post stuff that is extremely inappropriate but the only ones allowed in this group are adults and there is a pinned post and as well I welcome all the new members to the group with a message that specifically states that if they do find the posts here offensive then they are welcomed and encouraged to leave. Apparently we do have a troll in the group that we are attempting to ferret out but so far we have been unsuccessful..... But again, we are a closed and secret group and our posts cannot be shared to the everyday user in the facebook community."

RECOMMENDATIONS

Gone are the days when social media was used as a simple social networking tool. The technological features that platforms such as Twitter, Instagram, and Facebook now offer—like the ability to stream video in real-time—establish it as so much more. Social media is where we receive news; it's where we debate; it's where we organize. The fact that companies are responsible for moderating content on these platforms means they have the ability to curate the information we see, shape our opinions, and influence social movements.[49] Such an enormous responsibility warrants accountability to users. The following is a set of best practices for companies that seeks to provide concrete mechanisms for increasing accountability and transparency and improve user education. We hope that it can also serve as a set of recommendations for other civil society organizations to make in their advocacy for freedom of expression and transparency.

Practice Transparency

Transparency helps users better understand what constitutes a policy violation, how the internal takedown process works, and a path for remedy. For these reasons it's important for companies to build transparency and accountability mechanisms into their operations. They can do this in several ways; in this report we recommend two:

Expand Transparency Reporting

All of the companies that Onlinecensorship.org currently researches produce transparency reports.[50][51][52][53] These reports largely focus on data requests and content takedown requests issued by local or national law enforcement, not by other users.

Content removals made on the basis of user reporting under the companies' community guidelines or terms of service are not revealed. As privatized enforcement of hate speech and terrorism increases, there is a risk of further opacity, as government requests are increasingly re-directed through undisclosed "voluntary" terms of service takedowns. This lack of transparency hinders us from seeing the full picture. Free speech advocates, policymakers, researchers, and everyday users cannot advocate for progressive decisions

[49] Michael Nunez, "Former Facebook workers: we routinely suppressed conservative news," *Gizmodo*, 9 May, 2016. https://gizmodo.com/former-facebook-workers-we-routinely-suppressed-conser-1775461006

[50] Facebook, "Government Requests Report," https://govtrequests.facebook.com/.

[51] Twitter, "Twitter transparency report," https://transparency.twitter.com/en.html.

[52] Yahoo, "Transparency report: Overview," https://transparency.yahoo.com/.

[53] Google, "Google Transparency Report," https://www.google.com/transparencyreport/.

about different kinds of content without understanding the issues at scale.

Ranking Digital Rights, a global corporate accountability index for Internet and telecommunication companies, has recommended that companies "expand their transparency reporting to include requests from private parties as well as those from governments."[54]

"Without company reporting on private requests, the public has no insight into the influence of self-regulatory organizations such as the Internet Watch Foundation or the effect of policies related to copyright infringement, hate speech, the 'right to be forgotten,' ruling and other topics," the report reads.

We agree with this recommendation. In particular, we recommend providing details on the number of requests, breaking down the percentage of requests complied with by policy, country, and mechanism of takedown (whether flagging, filters, or other mechanisms).

Make the Content Moderation Process Transparent

Our user reports suggest widespread misinterpretations among users of how content moderation works. This stems from vague explanations of how violations are evaluated and a lack of detail in community guidelines, not to mention disparity in enforcement from country to country. Furthermore, there are many negotiations that happen behind the scenes concerning content moderation of which social media users are unaware—agreements between companies and governments, for example, can be dangerous since, as far as we are aware, policy makers lack insight into the scope and scale of the content that they are regulating. Companies are regulating both illegal speech and terms of service-banned speech, but users don't know which is governed by the state vs. a company's terms of service. In speaking with companies, we've found that they should:

- Clarify community guidelines, providing examples so that users can understand what constitutes a violation;
- Include messaging on the mechanisms used to evaluate content at multiple points of communication across the site and app: this may include community guidelines, in-app notifications of content takedowns and "report content" sections. Though many companies provide at least some of this information in various sections, the reports suggest users are having trouble finding and interpreting it. Having both a centralized location for policy information (as YouTube already does with its Policy Center)[55] and dispersed repetitions of this messaging will improve user education;
- Include a "more info" section in both in-app notifications of content takedowns and in the "report abusive content" box that explains the evaluation process and reason for the takedown;
- Indicate to users when they are notified of a content take down whether the content has been taken down because it was flagged by another user, or because of another reporting mechanism;
- Establish transparency toward civil society regarding government agreements. Clearly distinguish between different regulatory

[54] Priya Kumar, "Ranking Digital Rights Findings on Transparency Reporting and Companies' Terms of Service Enforcement," Ranking Digital Rights, 1 March 2016, https://rankingdigitalrights.org/wp-content/uploads/2016/03/RDR-Transparency-Findings.pdf.

[55] YouTube, "Policy Center," https://www.youtube.com/yt/policyandsafety/policy.html

systems: speech disallowed by law in specific jurisdictions versus speech disallowed by company Terms of Service.

- Detail how appeals are evaluated and educate users;

- Include content moderation and censorship guidelines in user education tools, such as Facebook's Security and Privacy Check Up wizards.[56]

Offer Redress

In cases where a user's content is removed in error, there must be an established system of redress where users can push back. Though more users report being aware of and utilizing appeals systems than in our last report, they remain distrustful of the appeals mechanisms available to them and expressed frustration at their perceived inability to get a human response. As it stands, the policy responses for erroneous content removals remain ad hoc, reactive, and offer little opportunity for users to get their content reinstated. We recommend that companies:

- Provide users information on appeals processes through multiple pages on the site. Though some users are receiving this information through in-app notifications, others appear to be missing these messages. Offering publicly available and prominently displayed information on what may be appealed and how appeals will be evaluated will improve users' perceptions that they do have opportunity for redress;

- Provide detailed language as to why an appeal was/was not restored, including explicit reference to the policy in question;

- Explain to users why certain forms of content cannot be appealed, where this is the case. For example, Facebook only allows appeals for account suspensions and page removals, and Instagram only allows appeals for account suspensions. Clarity on why individual posts cannot be appealed would give users a better understanding of the process. Moreover, where automatic bans are instituted, users are seeking a better understanding of why this form of enforcement was chosen as opposed to other mechanisms that allow redress.

For more information on companies' appeals procedures, see our resource on how to appeal.[57]

Encourage Best Practices Through User Education

In some cases, users' content was removed or they were blocked from a platform without knowing the specific reason. This reactive, punishment-based approach does more harm than good. Companies should educate their users about why they did wrong and citing which specific policies are being violated, instead of just taking them off the platform for long periods of time.

In the past year, we've seen a trend toward employing temporary bans for violations of community standards. While this may be an improvement in some ways from previous practices, these bans are punitive and don't effectively educate users about why what they did was wrong. Furthermore, they can potentially have devastating consequences for the user, who may be cut off from accessing the platform for work or using

[56] Facebook, "Security Checkup," https://www.facebook.com/help/443357099140264; Facebook, "Privacy Checkup," https://www.facebook.com/help/443357099140264.

[57] Onlinecensorship.org, "How to Appeal," https://onlinecensorship.org/resources/how-to-appeal.

[58] Jillian York, "Getting banned from Facebook can have unexpected and professionally devastating consequences," *Quartz*, 31 March 2016, http://qz.com/651001/t/371270/.

third-party services that require connection through the given service.[58]

There is little transparency as to how bans are applied to users who violate the rules. Users who have experienced such bans frequently report frustration with the process:

- One user said "I honestly actually have no idea what, who, or why I've been banned this time...Like I said, all I want is an explanation so that if nothing else I can try to avoid this in the future!"

- Another user said "I'm a professional photographer, and I had submitted a portfolio of images from a portrait series I'm working on to a French art photography website. They published the portfolio last Saturday, and I posted a link to the site. Facebook picks an image to go along with the link, and out of 12 images only two contained nudity--Facebook's algorithm chose one of those. I didn't know how to choose another image, and as I hadn't chosen it I didn't see any reason to be concerned, however, Facebook pulled the image down after about 8 hours, and suspended my account until I verified that none of the other images I'd posted over the years contained nudity, which I thought was odd, as their censors could've just looked at my images for themselves. I then posted the following in my Facebook timeline: Remember that post I made with the link to the French website that published my images? Well, Facebook pulled it down, claiming I'd posted nudity here. I didn't. I posted a link to the website, and Facebook's algorithm chose an image that contained nudity.... the word 'Kafkaesque' comes to mind."

Furthermore, some users have expressed confusion about why their content is removed and a desire to better understand the rules and how they're enforced. While several companies have robust help sections, this information is often disorganized and difficult to access, particularly on mobile devices. Companies should make it easier for users to access information about the community standards and their enforcement through the platform interface. A good example of how they might do this is Facebook's "Security Checkup" feature.[59] Companies should also consider implementing practices for obtaining forward user consent regarding content standards, particularly as they change.

Implement Responsible Policymaking

Social media companies are increasingly the gatekeepers of online speech. Companies that were once mere tech companies now have the ability to shape public opinion. As such, it's imperative that their policies surrounding content moderation are well-defined and publicly available. This isn't always easy; companies themselves have difficulty delineating their policies especially when every piece of content that is posted is nuanced and requires context for interpretation. A nude breast in a campaign to prevent breast cancer or a small nude child fleeing war is not the same thing as pornography.

Life happens in complex tones, and no one-size-fits-all content policy or automated algorithm can deliver the required sensitivity or context for what is truly in the public interest. Even still, there's more companies could do to ensure their policies ultimately decrease censorship and increase freedom of expression. Companies should:

- Develop a set of publicly stated values or principles to be weighed by content moderators as a mechanism for editorial transparency, including a statement of how the company defines the public interest. This statement could be augmented during particular conflicts or events to provide detail into the considerations that have shaped the company's policy line on certain kinds of content;

[59] Facebook Help Center, "What's Security Checkup and how do I start it?," https://www.facebook.com/help/799880743466869.

- Present users who are reporting content with a description of what types of speech violate community standards on the reporting page. Clear definitions for hate speech and harassment should also be included so that users can know whether the content they are reporting actually violates the guideline;

- Train content moderators to be globally-minded, and be transparent about where policies may be enforced differently in certain countries due to local legal restrictions or cultural norms;

- Review community guidelines for international awareness. Companies should ask themselves what global norms and values are being asserted in their guidelines regardless of local context.

ACKNOWLEDGEMENTS

The team would like to thank the Knight Foundation for their generous funding of Onlinecensorship.org through the Knight News Challenge. We would also like to thank Rebecca Jeschke, Jeremy Malcolm, and Corynne McSherry at EFF, and Ramzi Jaber at Visualizing Impact for their help in producing this report.