Who Moderates the Social Media Giants?

A Call to End Outsourcing

PAUL M. BARRETT

NYU | STERN

Center for Business
and Human Rights

June 2020

Contents

Acknowledgments

We extend special thanks to researchers and Stern Signature Project participants Abhinav Krishna, Tiffany Lin, and Su-Kyong Park.

Thanks also to the following for their time and insights:

Facebook: Monika Bickert, Ruchika Budhraja, Arun Chandra, Nick Clegg, Crystal Davis, Sarah Oh, Maxime Prades, Drew Pusateri, Guy Rosen, Miranda Sissons

Google/YouTube: Markham Erickson, Mike Grosack, Alex Joseph, Radha Penekelapati, Alexandria Walden, Clement Wolf

Twitter: Del Harvey, Nick Pickles

Other: Joshua Brustein of *Bloomberg News*; Sean Burke; Adelin Cai, formerly of Pinterest, Twitter, and Google; Cori Crider of Foxglove; Yael Eisenstat of Cornell Tech and formerly of Facebook; Debrynna Garrett; Christopher Gray; Sam Gregory of Witness; Jennifer Grygiel of Syracuse University; Clifford Jeudy; Sharon Kann of Media Matters for America; David Kaye of the United Nations and the Law School of the University of California, Irvine; Daphne Keller of Stanford Law School and formerly of Google; Jason Kint of Digital Content Next; Kate Klonick of St. John's Law School; Jay Lechner of Lechner Law; Roger McNamee, formerly of Elevation Partners; Casey Newton of *The Verge*; Britt Paris of Rutgers University; Tom Phillips, formerly of Google; Lawrence Pintak of Washington State University; Karen Rebelo of Boom Live; Sarah Roberts of the University of California, Los Angeles; Aaron Sharockman of PolitiFact; Diane Treanor of Coleman Legal Partners; Dave Willner of Airbnb and formerly of Facebook; Nicole Wong, formerly of Google and Twitter; Jillian York of the Electronic Frontier Foundation; Valera Zaicev.

This report was made possible by generous support from the John S. and James L. Knight Foundation and Craig Newmark Philanthropies.

Author
Paul M. Barrett is the Deputy Director of the New York University Stern Center for Business and Human Rights.

Executive Summary

Content moderation—the process of deciding what stays online and what gets taken down—is an indispensable aspect of the social media industry. Without it, online platforms would be inundated not just by spam, but by personal bullying, neo-Nazi screeds, terrorist beheadings, and child sexual abuse.

Despite the centrality of content moderation, however, major social media companies have marginalized the people who do this work, outsourcing the vast majority of it to third-party vendors. A close look at this situation reveals three main problems:

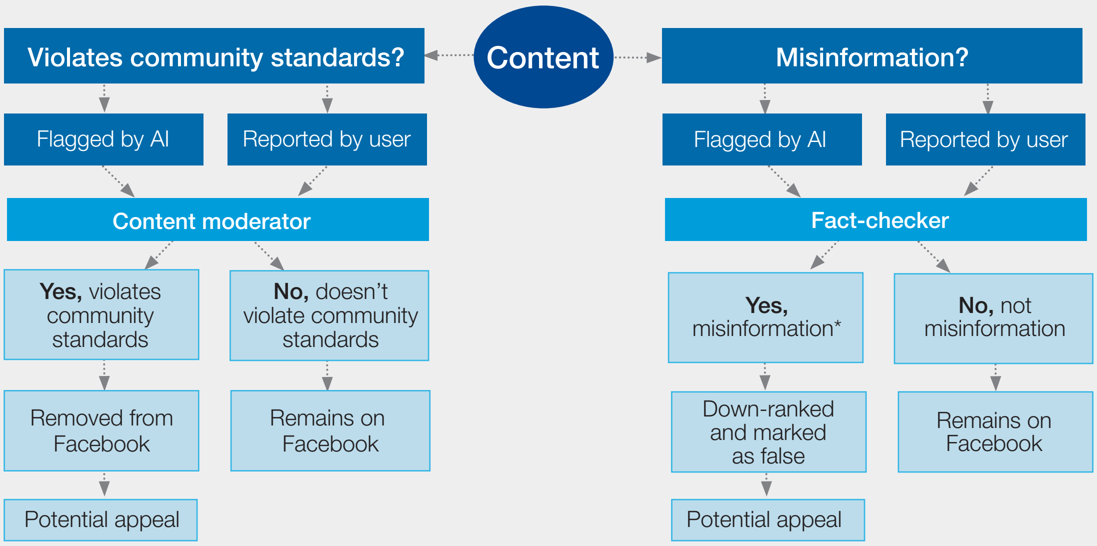
— In some parts of the world distant from Silicon Valley, the marginalization of content moderation has led to social media companies paying inadequate attention to how their platforms have been misused to stoke ethnic and religious violence. This has occurred in places ranging from Myanmar to Ethiopia. Facebook, for example, has expanded into far-flung markets, seeking to boost its user-growth numbers, without having sufficient moderators in place who understand local languages and cultures.

— The peripheral status of moderators undercuts their receiving adequate counseling and medical care for the psychological side effects of repeated exposure to toxic online content. Watching the worst social media has to offer leaves many moderators emotionally debilitated. Too often, they don't get the support or benefits they need and deserve.

— The frequently chaotic outsourced environments in which moderators work impinge on their decision-making. Disputes with quality-control reviewers consume time and attention and contribute to a rancorous atmosphere.

Outsourcing has become a common aspect of the globalized economy. Examples include customer-help centers in the Philippines, digital device factories in China, and clothing-production facilities in Bangladesh. Outsourcing is not inherently detrimental—if workers are paid fairly and treated humanely. A central question raised by outsourcing, in whatever industry it occurs, is whether it leads to worker exploitation. In social media, there's an additional concern about whether outsourcing jeopardizes optimal performance of a critical function.

How Facebook handles harmful content



*Misinformation is removed if it creates a threat of physical harm or interferes with voting or the Census.

(Executive Summary continued)

Today, 15,000 workers, the overwhelming majority of them employed by third-party vendors, police Facebook's main platform and its Instagram subsidiary. About 10,000 people scrutinize YouTube and other Google products. Twitter, a much smaller company, has about 1,500 moderators. These numbers may sound substantial, but given the daily volume of what is disseminated on these sites, they're grossly inadequate.

The enormous scale of the largest platforms explains why content moderation requires much more human effort to be done right. Every day, billions of posts and uploads appear on Facebook, Twitter, and YouTube. On Facebook alone, more than three million items are reported on a daily basis by users and artificial intelligence screening systems as potentially warranting removal. This degree of volume didn't happen by accident; it stems directly from Facebook's business strategy of relentlessly pursuing user growth in an effort to please investors and the advertisers that are the company's paying customers.

Focusing primarily on Facebook as a case study, this report begins with an Introduction and overview, followed by a sidebar on page 6 about the interplay between the coronavirus pandemic and content moderation. Part 2 describes the origin and early development of moderation. Infographics providing a statistical breakdown of content moderation by Facebook, YouTube, and Twitter appear on pages 10 and 11.

Part 3 examines problems with Facebook's content moderation, with an emphasis on the lack of adequate health care for the people who do it and the generally chaotic environment in which they work. This section of the report draws heavily on interviews with former moderators who describe meager "wellness" programs in workplaces characterized by a surprising degree of contentiousness.

Part 4 looks at the lack of adequate moderation in at-risk countries in regions such as South Asia. A sidebar on fact-checking, a variant of content moderation, appears on page 23. And finally, Part 5 offers recommendations for improving the situation, which we also provide here, in capsule form:

Summary of our Recommendations

1 **End outsourcing of content moderators and raise their station in the workplace.** Facebook—and YouTube and Twitter—should gradually bring on board, with suitable salaries and benefits, a significant staff of content moderators drawn from the existing corps of outsourced workers and others who want to compete for these improved jobs.

2 **Double the number of moderators to improve the quality of content review.** Members of an expanded moderator workforce would have more time to consider difficult content decisions while still making these calls promptly.

3 **Hire a content overseer.** To streamline and centralize oversight, Facebook—and the other platforms—each should appoint a senior official to supervise the policies and execution of content moderation.

4 **Expand moderation in at-risk countries in Asia, Africa, and elsewhere.** The citizens of these nations deserve sufficient teams of moderators who know local languages and cultures—and are full-time employees of the social media companies.

5 **Provide all moderators with top-quality, on-site medical care.** At the center of this improved medical care should be the question of whether a given employee is capable of continuing to moderate the most disturbing content.

6 **Sponsor research into the health risks of content moderation.** While the danger of post-traumatic stress disorder and related conditions seems obvious, the social media companies still lack a sufficient understanding of the precise risks their moderators face. High-quality academic research is needed to address this gap.

7 **Explore narrowly tailored government regulation.** One interesting idea comes from Facebook itself. The company suggests government oversight of the "prevalence" of harmful content, which it defines as the frequency with which deleterious material is viewed, even after moderators have tried to weed it out.

8 **Significantly expand fact-checking to debunk mis- and disinformation.** Disproving conspiracy theories, hoaxes, and politically motivated mis- and disinformation is a noble pursuit but one that's now being done on too small a scale.

1. Introduction

“‘Content moderators are the people literally holding this platform together. They are the ones keeping the platform safe.’
— A Facebook design engineer participating on an internal company message board, 2019”

Picture what social media sites would look like without anyone removing the most egregious content posted by users. In short order, Facebook, Twitter, and YouTube (owned by Google) would be inundated not just by spam, but by personal bullying, neo-Nazi screeds, terrorist beheadings, and child sexual abuse. Witnessing this mayhem, most users would flee, advertisers right behind them. The mainstream social media business would grind to a halt.

Content moderation—deciding what stays online and what gets taken down—is an indispensable aspect of the social media industry. Along with the communication tools and user networks the platforms provide, content moderation is one of the fundamental services social media offers—perhaps the most fundamental. Without it, the industry’s highly lucrative business model, which involves selling advertisers access to the attention of targeted groups of users, just wouldn’t work.[1]

“Content moderators are the people literally holding this platform together,” a Facebook design engineer reportedly said on an internal company message board during a discussion of moderator grievances in early 2019. “They are the ones keeping the platform safe. They are the people Zuck [founder and CEO Mark Zuckerberg] keeps mentioning publicly when we talk about hiring thousands of people to protect the platform.”[2]

And yet, the social media companies have made the striking decision to marginalize the people who do content moderation, outsourcing the vast majority of this critical function to third-party vendors—the kind of companies that run customer-service call centers and back-office billing systems. Some of these vendors operate in the U.S., others in such places as the Philippines, India, Ireland, Portugal, Spain, Germany, Latvia, and Kenya. They hire relatively low-paid labor to sit in front of computer workstations and sort acceptable content from unacceptable.

The coronavirus pandemic has shed some rare light on these arrangements. As the health crisis intensified in March 2020, Facebook, YouTube, and Twitter confronted a logistical problem: Like millions of other workers, content moderators were sent home to limit exposure to the virus. But the platforms feared that allowing content review to be done remotely from moderators’ homes could lead to security and privacy breaches. So the social media companies decided temporarily to sideline their human moderators and rely more heavily on automated screening systems to identify and remove harmful content. In normal times, these systems, powered by artificial intelligence (AI), identify and, in some cases, even eliminate, certain disfavored categories of content, such as spam and nudity. Other categories, including hate speech and harassment, typically still require human discernment of context and nuance.

In unusual public concessions, Facebook, YouTube, and Twitter acknowledged that depending more on automation would come at a price: Used on their own, the AI systems are prone to removing too much content. The algorithms "can sometimes lack the context that our teams [of human moderators] bring, and this may result in us making mistakes," Twitter said.[3] These admissions underscored the continuing fallibility of automated screening and the corresponding value of human review. As of this writing, outsourced human moderation was just beginning to resume, as some Facebook reviewers returned to their offices on a voluntary basis and others were allowed to do certain work from home. (For more on the coronavirus and moderation, please see the sidebar on page 6.)

In more ordinary times, three other problems arise from the outsourcing of content moderation. First, in some parts of the world distant from Silicon Valley, the marginalization of moderation has led to the social media companies paying inadequate attention to how their platforms have been misused to stoke ethnic and religious violence. This has occurred in places such as Myanmar and Ethiopia. Facebook, for example, has expanded into far-flung markets, seeking to boost its user-growth numbers, without having sufficient moderators in place who understand local languages and cultures. Second, the peripheral status of moderators undercuts their receiving adequate counseling and medical care for the psychological side effects of repeated exposure to toxic online content. Watching the worst social media has to offer leaves many moderators emotionally debilitated. Too often, they don't get the support or benefits they need and deserve. And third, the frequently chaotic outsourced environments in which moderators work impinge on their decision-making.

Despite the crucial role they perform, content moderators are treated, at best, as second-class citizens. Some full-time employees recognize the anomaly. "Why do we contract out work that's obviously vital to the health of this company and the products we build?" a Facebook product manager asked during the same early-2019 in-house message board exchange about moderators' discontent.[4]

'Plausible Deniability'

According to Sarah Roberts, a pioneering scholar of content moderation, the social media companies handle the activity in a fashion that diminishes its importance and obscures how it works. "It's a way to achieve plausible deniability," Roberts, an information studies expert at the University of California, Los Angeles, says in an interview. "It's a mission-critical function, but you fulfill it with your most precarious employees, who technically aren't even your employees."[5]

Beyond the distancing effect identified by Professor Roberts, outsourcing saves social media companies significant amounts of money on moderation, just as it lowers costs for janitorial, food, and security services. Contract moderators don't enjoy the generous pay scales and benefits characteristic of Silicon Valley. Outsourcing also has given the tech companies greater flexibility to hire moderators in a hurry without having to worry about laying them off if demand for their services wanes.

Similar factors have motivated outsourcing by Western corporations in other contexts, ranging from customer-help centers in the Philippines to digital device factories in China and clothing-production facilities in Bangladesh. Outsourcing, to be sure, is not inherently detrimental. Bangladeshi women stitching t-shirts and trousers for sale in the U.S. and Europe need employment and may not have better options. If workers are paid fairly and treated humanely, outsourcing can represent a salutary aspect of the globalized economy. Likewise, social media moderators may take the job eagerly, however modest the wage or harrowing the subject matter.

A central question raised by outsourcing, in whatever industry it occurs, is whether it leads to worker exploitation. In social media, there's an additional concern about whether outsourcing jeopardizes optimal performance of a critical function.

Each of the social media platforms began to do content moderation on a limited basis soon after the start of its operations—Facebook in 2004, YouTube in 2005, and Twitter in 2006. Improvising at first, employees fielded user complaints about offensive posts or comments the way they might deal with forgotten passwords. The process has evolved since then, as the platforms have adopted elaborate sets of standards to guide moderation and built automated screening systems relying on AI. Moderation has become a hybrid of algorithmic and human analysis in which live reviewers assess huge volumes of potentially problematic posts spotted by users or AI technology.

Today, 15,000 workers, the overwhelming majority of them employed by third-party vendors, police Facebook's main platform and its Instagram subsidiary. About 10,000 people scrutinize YouTube and other Google products. Twitter, a much smaller company, has about 1,500 moderators. These numbers may sound substantial, but they're woefully inadequate, Jennifer Grygiel, a social media scholar at Syracuse University, says in an interview. "To get safer social media, you need a lot more people doing moderation."[6]

A Problem of Scale

The enormous scale of the largest social media platforms explains why content moderation requires additional human effort. Every day, billions of posts and uploads appear on Facebook, Twitter, and YouTube. On Facebook alone, more than three million items are reported on a daily basis by AI systems and users as potentially warranting removal.[7] This degree of volume didn't happen by accident; it stems directly from Facebook's business strategy of relentlessly pursuing user growth in an effort to please investors and the advertisers that are the company's paying customers.

A moderation workload of this heft, combined with the complexity of some of the decisions, naturally leads to errors. Zuckerberg conceded in a November 2018 white paper that moderators "make the wrong call in more than one out of every 10 cases." He didn't specify how many erroneous calls that equates to, but if Facebook moderators review three million posts a day, Zuckerberg's 10% error rate implies 300,000 blunders every 24 hours. To his credit, the CEO admitted that given the size of Facebook's user base—now some 2.5 billion people—"even if we were able to reduce errors to one in 100, that would still be a very large number of mistakes."[8]

The gargantuan volume of online material creates other problems, as well, including the difficulty of devising general moderation rules that encompass the circumstances of billions of daily posts and uploads. In a landmark episode from 2016, Facebook removed a frontal photo image of a naked nine-year-old Vietnamese girl running in terror from an explosion. A straightforward violation of the platform's rules against nudity and child exploitation? Not necessarily. The 1972 photo, informally known as "Napalm Girl," had won a Pulitzer Prize and remains an iconic representation of the Vietnam War. After a public outcry, Facebook reversed itself, restored the picture, and created an exception for otherwise offensive content that's "newsworthy."[9]

The "Napalm Girl" incident illustrates that content moderation will never achieve perfect results or please everyone. Some harmful content will persist on mainstream platforms, and moderators will sometimes censor items unwisely. But these realities should not become an argument against incremental improvement, which this report urges in the recommendations that begin on page 24.

Facebook does a variation of content review called third-party fact-checking, which also deserves consideration. Fact-checking evaluates not whether content falls into a forbidden category like hate speech, but instead whether content is true or false. Facebook outsources fact-checking but in a different way from how it handles content moderation and in a fashion that we support. The company hires journalism organizations and specialty fact-checking websites to determine whether content flagged by users or AI is accurate. When it's identified, false content typically is not removed. Facebook labels and down-ranks it in users' News Feed so that fewer people see it. Scale becomes an issue for fact-checking, as it does for content moderation. Facebook sends flagged content to more than 60 fact-checking organizations worldwide, but each organization typically assigns only a handful of reporters to investigate Facebook posts. The number of potentially false Facebook items far exceeds fact-checkers' capacity, meaning that the system is overwhelmed and ultimately inadequate to the task. Like content moderation, fact-checking demands more resources. (For more on fact-checking, please see the sidebar on page 23 and our recommendations on page 24.)

In an era when much of our political and cultural expression takes place online, content moderation and fact-checking, while little understood by the broader public, play an important role helping to shape democracy and society at large. Social media companies could improve their performance by bringing content review closer to the core of their corporate activities, greatly increasing the number of human moderators (even while continuing to refine AI screening software), and elevating moderators' status to match the significance of their work. On the fact-checking front, Facebook ought to use its ample resources to expand capacity as well, while Twitter and YouTube, however belatedly, need to follow Facebook's example. Given the stakes for social media users, and everyone else affected by what happens online, the companies have an obligation to take swift, dramatic action.

Why Focus on Facebook?

All three of the major social media platforms—as well as many others—undertake content moderation.
The following pages, however, focus primarily on Facebook. There are three reasons for a case study approach: First, Facebook, headquartered in Menlo Park, Calif., deserves close scrutiny because it's the largest competitor in its segment of the industry and has served as a trend-setter in content moderation. Second, putting one company under the microscope allows for a more detailed look at corporate practices. From this examination, lessons can be developed and applied to other companies, as well. And third, Facebook was more forthcoming than its rivals YouTube and Twitter. That we have rewarded this responsiveness with more in-depth attention may strike some people at Facebook as ironic, if not downright irritating. We hope that upon reading the report, they will find it tough-minded but fair.

Facebook's openness, we should emphasize, went only so far. Like YouTube and Twitter, Facebook turned down our repeated requests to visit one or more of its moderation sites. This denied us access to current reviewers and obliged us to seek out individuals who formerly did moderation work. More broadly, Facebook declined to answer a number of our questions, including basic ones such as what percentage of its moderator workforce it outsources. Still, we would like to think that Facebook's greater communicativeness overall indicates a willingness to consider our recommendations—and serve as an example to other platform companies, which bear the same responsibility as Facebook to improve how they do content moderation.

The Coronavirus Pandemic and Content Moderation

The coronavirus pandemic has shaken the global economy to its foundation, causing factory closures, transportation shut-downs, and mass layoffs. A more modest effect in the social media industry concerned content moderation. In the name of social distancing, thousands of reviewers were sent home. But as noted in the main text of this report, Facebook, YouTube, and Twitter didn't want content review to take place remotely for fear of potential software-security breaches or user-privacy violations. All three social media companies announced in mid-March that they would temporarily reduce their reliance on human moderation and shift more of the content-review burden to their AI-driven technology.[1]

Facebook went a step further. In light of the stay-at-home edicts affecting many of the company's outsourced moderators, Mark Zuckerberg explained during a March 18, 2020, teleconference with reporters that he had decided to enlist some of the company's full-time employees to handle the review of "the most sensitive types of content." He mentioned content related to suicide and self-harm, child exploitation, and terrorist propaganda. As with the shift to more reliance on AI, Zuckerberg wasn't specific about how long the hand-off of responsibility for sensitive content would last. Facebook would stick with the rearrangement, he said, "for the time being."[2]

By early May, a small number of Facebook moderators were beginning to return to their workstations at offices run by third-party vendors. As the grip of the pandemic loosens, all three of the major social media companies are expected to reestablish the outsourcing structure that existed before the health crisis. Part of the reason for this is that AI can't get the job done on its own.[3]

In his session with journalists, Zuckerberg conceded that the pandemic-induced reliance on technology would lead to more mistakes. The company's algorithms inevitably would "take down some content that was not supposed to be taken down," he said. This prediction soon proved correct. In one illustration, Facebook's automated system—calibrated to ban virus profiteering—mistakenly flagged posts by volunteers making protective masks for doctors and nurses.[4] YouTube and Twitter made similar statements warning of overly aggressive algorithms. All of these unusual concessions about the shortcomings of technology strongly imply a continuing need for human involvement in content moderation.

Another lesson from the pandemic is that more of this human involvement should come from people who are full-time Facebook employees—like the full-timers given responsibility for reviewing sensitive content during the coronavirus emergency. Facebook should use its response to the public health calamity as a pilot project to assess the feasibility of making all content moderators Facebook employees. The sense of trust that Zuckerberg has in his own people—signaled by his turning to them during a time of national crisis—suggests an opening for discussion within Facebook's senior ranks.

Central to our recommendations, which begin on page 24, is the idea that, quite apart from temporary pandemic work-arounds, Facebook and its rival platforms need to end the outsourcing of responsibility for content review. Doing so would address the trio of dangers that outsourcing creates: first, that at-risk countries receive insufficient attention from moderators; second, that moderators' mental health is not adequately protected; and third, that the outsourced environment is not conducive to the sort of careful content assessment that's vital for user safety.

[1] Elizabeth Dwoskin and Nitasha Tiku, "Facebook Sent Home Thousands of Human Moderators Due to the Coronavirus. Now the Algorithms Are in Charge," *The Washington Post*, March 23, 2020, https://www.washingtonpost.com/technology/2020/03/23/facebook-moderators-coronavirus/.

[2] Mark Zuckerberg, "Media Call," Facebook, March 18, 2020, https://about.fb.com/wp-content/uploads/2020/03/March-18-2020-Press-Call-Transcript.pdf.

[3] "Coronavirus: Facebook Reopens Some Moderation Centers," *BBC*, April 30, 2020, https://www.bbc.com/news/technology-52491123.

[4] Mike Isaac, "In a Crackdown on Scams, Facebook Also Hampers Volunteer Efforts," *The New York Times*, April 6, 2020, https://www.nytimes.com/2020/04/05/technology/coronavirus-facebook-masks.html.

2. The Origins and Development of Content Moderation

“‘We were supposed to delete things like Hitler and naked people,’ recalls Dave Willner, a member of Facebook’s pioneering content moderation group. He and his colleagues were told to remove material ‘that made you feel bad in your stomach.’”

Begun haphazardly and developed on the fly, content moderation initially was intended to insulate users from pornography and intolerance. This aim has persisted, even as moderation also became a shield with which social media platforms have sought to fend off controversy and negative publicity, says Professor Roberts of UCLA. “It’s the ultimate brand-protection scheme. It’s brand protection in the eyes of users and in the eyes of advertisers.”

When Dave Willner arrived at Facebook in 2008, not long after earning his undergraduate degree in anthropology from Bowdoin College, content moderation was still a modest affair, performed in-house. It had its roots in the early years of the internet, when volunteers helped oversee chat rooms by reporting “offensive” content. During his second year at Facebook, Willner joined a team of 12 people who followed a list of moderation rules contained on just a single page. “We were supposed to delete things like Hitler and naked people,” he recalls in an interview. More generally, they removed content “that made you feel bad in your stomach.” Even though Facebook already had about 100 million users, the dozen in-house moderators didn’t feel overwhelmed, he says. With a primarily American clientele, Facebook “was still something mostly for college students and recent graduates, and most of us were recent graduates, so for the most part, we understood what we were looking at.” Early in their corporate lives, YouTube and Twitter gave their moderators similarly bare-bones instructions of what to remove.

Willner and others involved in what might be called the artisanal phase of content moderation could do their jobs without fear of legal repercussions, at least within the U.S. This was thanks to a federal provision known as Section 230 of the Communications Decency Act of 1996. An extraordinary boon to online commerce, the law shields internet platforms from liability for most content posted by users. This protection applies even if platforms actively moderate user-generated content. According to St. John’s University legal scholar Kate Klonick, “the existence of Section 230 and its interpretation by courts have been essential to the development of the internet as we know it today.”[10]

Seeking Simplicity

Empowered by Section 230, Willner took the lead in replacing Facebook’s one page of moderation rules with a more fully developed set of guidelines. The 15,000-word document he eventually produced remains the basis of the company’s publicly available Community Standards, which have been amended many times over the years. The standards favor free speech when possible, Willner says. But their overarching goal is to provide moderators with simple, concrete rules that can be applied consistently by nonlawyers. This aim became increasingly important as Facebook expanded, and content moderation expanded along with it.

By 2010, explosive growth at Facebook made it impossible for its in-house content-review team to handle the increased volume of user reports about spam, pornography, hatred, and violence. The social network needed more moderators. "There wasn't much debate about what to do, because it seemed obvious: We needed to move this to outsourcing," Willner says. "It was strictly a business-ops decision," based on cost concerns and the greater flexibility outsourcing offered. By 2013, when Willner left the company, he says, Facebook had more than a billion users and about 1,000 moderators, most of them now outsourced. This produced a ratio of one moderator for every million users.

Content moderation was outsourced in another sense, as well. Facebook relied heavily on users, acting without compensation, to report potentially offensive or dangerous content to the company. This reliance on what amounted to an enormous volunteer corps of harmful content scouts meshed with yet another aspect of the social media business model: the expectation that users would supply most of the material—ranging from puppy pictures to political punditry —that draws people to social media. These several forms of outsourcing have combined to help Facebook keep its full-time employee headcount—now at 45,000—considerably lower than it otherwise would be, while raising its enviable profit margins. Moreover, while the move to lower-cost outsourcing of content moderation might seem to some within Facebook as having been inevitable, it was, in fact, a purposeful choice, driven by financial and logistical priorities. Over time, this choice would have distinct consequences.

As Facebook grew, disturbing content on the site proliferated. In 2013, a public controversy flared when Facebook moderators failed to remove content from groups and pages featuring supposedly humorous references to sexual assaults on women. One of the "jokes" included the punch line, "Tape her and rape her," written over a photo of a woman with heavy white tape covering her mouth. Feminists expressed outrage, and certain large corporations threatened to pull their advertising from the platform. After initially insisting that the misogynistic material didn't violate its hate speech rules, Facebook reversed itself, announced that the rape jokes did breach its standards after all, and took them down.[11]

The clash over rape jokes illustrated Facebook's struggle to balance free expression against freedom from hatred and cruelty. Twitter leaned more decisively toward allowing users to speak freely. Unfortunately, this strategy made Twitter a notorious destination for trolls and hate groups, one where instances of harassment were, and to some extent still are, common.[12] In 2015, *Guardian* columnist Lindy West recounted her years-long experience with bullies who tweeted vile taunts at her, such as, "No one would want to rape that fat, disgusting mess." In response, Twitter's then-CEO, Dick Costolo, wrote a scathing internal memo, which promptly leaked. "We suck at dealing with abuse and trolls on the platform," Costolo wrote, "and we've sucked at it for years."[13]

Despite the challenges it presented for all of the platforms, content moderation continued to expand rapidly. By 2016, Facebook had 1.7 billion users and 4,500 moderators, most of them employed by outside contractors. The ratio of moderators to users was much improved from three years earlier but still stood at one to 377,777. Today, the ratio is one to 160,000, and the company uses far more automation to complement human reviewers.

Two key characteristics have reinforced the marginalization of the function. First, it's a source of almost exclusively bad news: Tech journalists and the public typically focus on content moderation when it fails or sparks contention, not on the countless occasions when it works properly. "No one says, 'Let's write a lengthy story on all of the things that didn't happen on Twitter because of successful moderation,'" observes Del Harvey, the platform's vice president for trust and safety.

Tom Phillips, a former executive at Google who left that company in 2009, makes a related point. Moderation has never been fully accepted into Silicon Valley's vaunted engineering-and-marketing culture, he says. "There's no place in that culture for content moderation. It's just too nitty-gritty."

Content moderation presents truly difficult challenges. Before the 2000s, corporations hadn't confronted a task quite like it. But the difficulty stems directly from the business models chosen by Facebook, YouTube, Twitter, and some of their smaller rivals. These models emphasize an unremitting drive to add users and demonstrate growth to investors. More users attract more revenue-generating advertising, but they also produce more content to moderate and more permutations of meaning, context, and nuance—all of which invite error.

A CEO's Concession

In the wake of the 2016 presidential election debacle, in which Russian operatives used Facebook, Instagram, and Twitter to spread disinformation, Mark Zuckerberg was on the defensive. Facebook wasn't keeping up with the content challenges it faced, he conceded in a February 2017 public essay: "In the last year, the complexity of the issues we've seen has outstripped our existing processes for governing the community." Moderators, he continued, were "misclassifying hate speech in political debates in both directions—taking down accounts and content that should be left up and leaving up content that was hateful and should be taken down." He pointed to the precipitous removal of "newsworthy videos related to Black Lives Matter and police violence"—content that often included raw language about race, but was uploaded in the spirit of combating racism. In many instances, the CEO added, the company's reliance on users to report troubling content simply wasn't working. He tried to deflect the blame for this. "We review content once it is reported to us," he wrote. "There have been terribly tragic events—like suicides, some live streamed—

that perhaps could have been prevented if someone had realized what was happening and reported them sooner. There are cases of bullying and harassment every day that our team must be alerted to before we can help out."[14]

Zuckerberg's suggestion that users bear primary responsibility for policing Facebook obscures that he and his business colleagues designed the system, flaws and all, and failed to anticipate how much harmful content Facebook would attract.

Three months later, Zuckerberg announced that he would increase the number of content moderators by two-thirds, to 7,500. When describing such expansions, the CEO and other Facebook executives generally haven't mentioned that the majority of moderators are outsourced workers, not full-time company employees. The 3,000 new moderator hires in 2017 followed public outcry over the posting of violent videos. One showed the fatal shooting of a 74-year-old retiree in Cleveland; another, a live stream, depicted a man in Thailand killing his 11-month-old daughter. It took Facebook moderators more than two hours to remove the Cleveland video. The footage from Thailand stayed up for about 24 hours and was viewed roughly 370,000 times. "If we're going to build a safe community, we need to respond quickly," Zuckerberg wrote in a post.[15]

In December 2017, *ProPublica* took a revealing look at moderation. The non-profit investigative journalism organization solicited from its readers instances where they believed Facebook had erred in applying its own standards. Of the more than 900 posts submitted, *ProPublica* asked Facebook to explain a sample of 49 items, most of which involved leaving up material that *ProPublica* and its readers perceived as hate speech. Facebook acknowledged that in 22 cases, its moderators had made mistakes. Even for a sample selected in this non-representative fashion, the 45% error rate was remarkable. "We're sorry for the mistakes we have made—they do not reflect the community we want to build," Facebook said in a statement at the time. "We must do better."[16]

(The main text continues on page 12.)

Ad Hoc Policies: From COVID-19 to Holocaust Denial

Content moderation is a large and complicated topic. This report covers moderation problems related to outsourcing but not to other dimensions of the subject. For example, it does not delve into questions surrounding high-level policy decisions about moderation made by senior executives. One illustration of such a determination was Facebook's laudable decision beginning in the winter of 2020 to act more aggressively than usual to remove dangerous misinformation related to the COVID-19 pandemic: false cures, conspiracy theories, and the like. Another high-level policy call of a very different sort concerns posts that deny that the Holocaust took place. Facebook continues, unwisely, to allow content that promotes the idea that the Holocaust never occurred, despite the fact that such content represents rank anti-Semitism.[1]

These policies, for better or worse, help determine the kind of material available to users on Facebook. But the decisions behind them presumably would be made regardless of whether content moderators work on an outsourced basis. An analogous decision in the fact-checking realm was Mark Zuckerberg's determination in the fall of 2019 that Facebook, in the name of free speech, would not review political advertising for untrue statements. Again, this was a weighty judgment call—one that unfortunately extended a virtual invitation to politicians and their campaigns to lie—but not one that bears on how Facebook structures its relationship with fact-checkers.[2]

Facebook has a serious and systematic process for routinely amending its Community Standards. But too often, high-level content policy decisions seem ad hoc and reactive. The company's decision-making about white nationalism and white separatism illuminates the problem. Before March 2019, Facebook had prohibited hateful attacks on people based on characteristics such as race and ethnicity. This prohibition included expressions of white supremacy. But the platform distinguished between white supremacy, on the one hand, and white nationalism and separatism, on the other. The distinction was based on the dubious notion that the latter could be bound up with legitimate aspects of people's identity.[3]

In 2018, the tech news site *Motherboard* published internal documents used to train Facebook content moderators. The documents showed that Facebook allowed "praise, support, and representation" of white nationalism and separatism "as an ideology." This sparked a new round of criticism from civil rights advocates, who had long contended that white nationalism and separatism stood for the same repugnant ideas as white supremacy. Under pressure from these advocates, Facebook changed its position, ultimately reaching the right result, although belatedly and in a roundabout manner.[4]

[1] Ezra Klein, "The Controversy Over Mark Zuckerberg's Comments on Holocaust Denial, Explained," *Vox*, July 20, 2018, https://www.vox.com/explainers/2018/7/20/17590694/mark-zuckerberg-facebook-holocaust-denial-recode; Jonathan A. Greenblatt, "Facebook Should Ban Holocaust Denial to Mark 75th Anniversary of Auschwitz Liberation," *USA Today*, January 26, 2020, https://www.usatoday.com/story/opinion/2020/01/26/auschwitz-liberation-ban-holocaust-denial-on-facebook-column/4555483002/.

[2] "Facebook's Zuckerberg Grilled Over Ad Fact-Checking Policy," *BBC*, October 24, 2019, https://www.bbc.com/news/technology-50152062.

[3] Tony Romm and Elizabeth Dwoskin, "Facebook Says It Will Now Block White-Nationalist, White-Separatist Posts," *The Washington Post*, March 27, 2019, https://www.washingtonpost.com/technology/2019/03/27/facebook-says-it-will-now-block-white-nationalist-white-separatist-posts/.

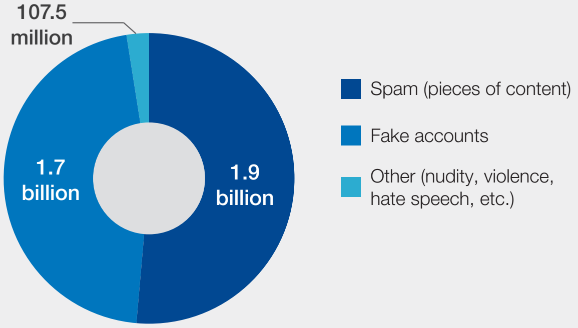
[4] Joseph Cox, "Leaked Documents Show Facebook's Post-Charlottesville Reckoning With American Nazis," *Motherboard*, May 25, 2018, https://www.vice.com/en_us/article/mbkbbq/facebook-charlottesville-leaked-documents-american-nazis.

Moderation by the numbers: removing harmful content

Beyond spam and fake accounts, Facebook contends with adult nudity, violence, and dangerous organizations...

Facebook content removed or covered*

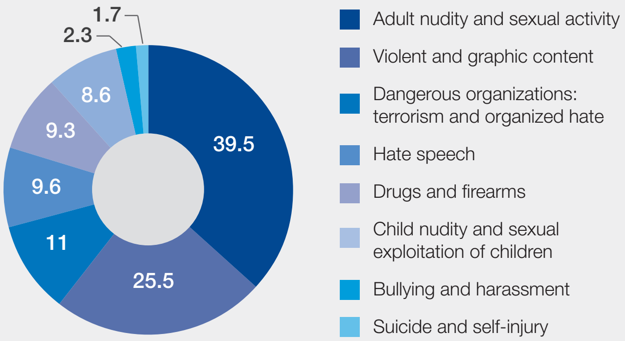
These figures are for the first quarter of 2020, the most recent available data.



*Facebook covers some nonviolating content and provides a warning that it may be disturbing.

Facebook removals other than fake accounts and spam*

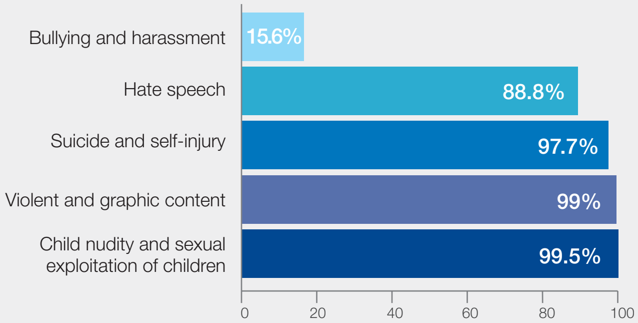
First quarter of 2020, in millions.



*Includes some content that is covered but not removed.

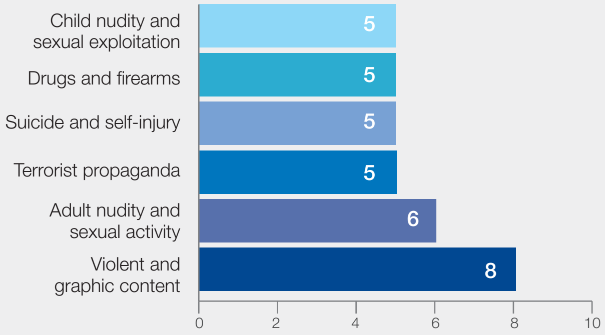
Heavy reliance on artificial intelligence

Percentage of content removed or covered that was flagged by Facebook AI technology before any users reported it (first quarter of 2020).



Prevalence of selected forms of harmful content

Prevalence measures how often harmful content slips past moderation efforts and remains available to users. This chart estimates the upper limit of views per 10,000 of content that violated the community standard in question.*



*Facebook generates prevalence estimates based on its own sampling of content.

Detailed rules govern content moderation

Facebook provides highly specific guidelines for moderators to enforce. Here's one example:

Hate speech

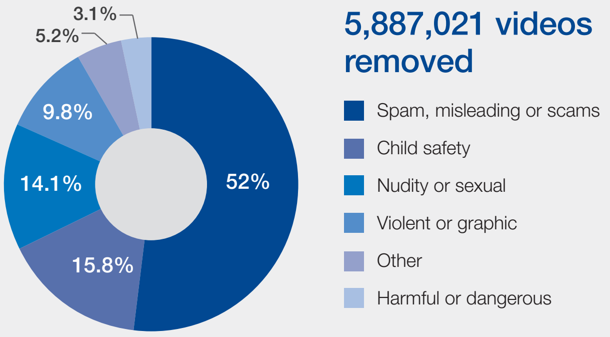
Facebook bans "direct attacks" on people based on "protected characteristics," such as race, ethnicity, national origin, religious affiliation, sexual orientation, gender, or disability. Direct attacks can take the form of "violent speech," "dehumanizing speech or imagery," and derogatory comparisons to insects, animals perceived as intellectually or physically inferior, filth, bacteria, disease, or feces. Certain designated comparisons are also prohibited, including Blacks and apes, Jews and rats, Muslims and pigs, and Mexicans and worm-like creatures.

Source: Facebook

...while YouTube takes down videos endangering children, and Twitter suspends accounts for hateful conduct.

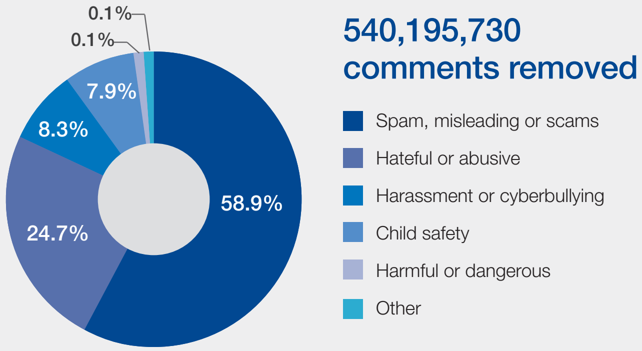
YouTube videos removed

These figures are for the fourth quarter of 2019, the most recent available data.



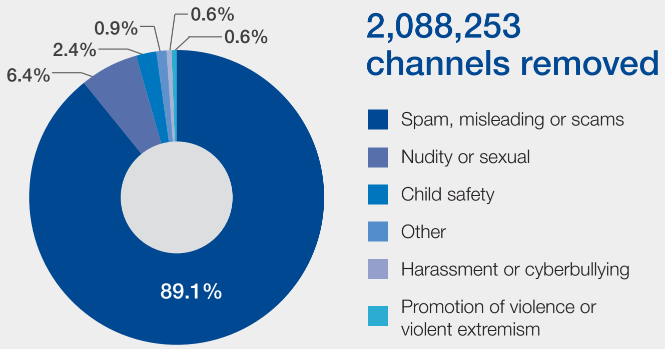
YouTube comments removed

Fourth quarter of 2019



YouTube channels removed

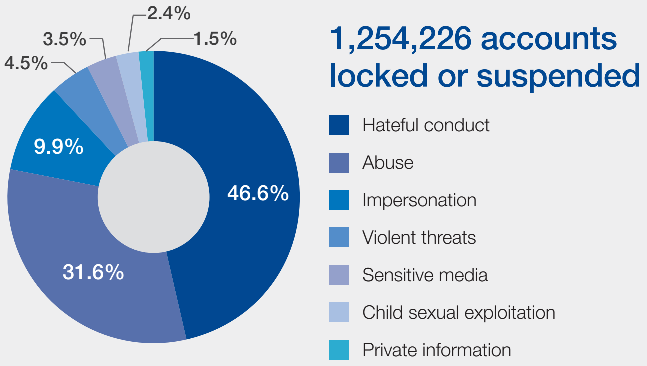
Fourth quarter of 2019



Source: YouTube

Twitter accounts locked or suspended*

These figures are for the first half of 2019, the most recent available data.



*Twitter enforcement actions range from temporarily disabling accounts to shutting them down altogether.

Selected Twitter enforcement statistics

First half of 2019

50% of tweets that Twitter took action on for abuse were proactively identified using technology, rather than being reported by users. This compares to 20% a year earlier.

105% more accounts overall were locked or suspended for violating rules.

119% more accounts were suspended for violating private information rules.

133% more accounts were locked or suspended for hateful conduct.

30% fewer accounts were suspended for promotion of terrorism.

Source: Twitter

3. The Moderator's Experience

> “‘We still have enforcement problems. You're going to hear that across the industry.’ — Monika Bickert, Facebook's vice president for global policy management”

Mark Zuckerberg once said that “in a lot of ways, Facebook is more like a government than a traditional company. We have this large community of people, and more than other technology companies we're really setting policies.”[17]

If Facebook is like a government, then Zuckerberg heads the executive branch, or perhaps rules as monarch. The legislature takes the form of a policy team of more than 100 people working under a company vice president named Monika Bickert. A former federal prosecutor, Bickert leads a rolling process of supplementing and amending Facebook's public Community Standards, as well as its internal guidelines interpreting the standards. Content moderators, in this scheme, are the police officers who enforce the standards. To do so, the cops on the beat rely on leads from two types of informants: human users and, increasingly, inanimate AI flagging systems.

The moderators don't work for the Facebook government, however. They are rent-a-cops, employed by third-party vendors.

By all accounts, Bickert's policy group works with great earnestness to craft highly granular rules that she says are intended to remove as much discretion as possible from moderators. Specificity “makes it possible for us to apply policies at this unprecedented scale of billions of posts every day,” she adds. But asked about Zuckerberg's estimate of a 10% error rate, she bluntly acknowledges: “We still have enforcement problems. You're going to hear that across the industry.”

To round out Zuckerberg's government metaphor, Facebook has launched the equivalent of a judicial branch. A nascent semi-autonomous Oversight Board—informally referred to as the Supreme Court of Facebook—will be populated by 40 outsiders. Facebook named the first 20 in May 2020, and the list included an impressive, diverse array of legal scholars, human rights experts, and former public officials. Working in panels of five, board members will review selected user appeals from moderators' decisions. Apart from specific cases, the board will respond to requests for policy guidance from Facebook. These activities are designed to resolve particular disputes and establish principles that will guide the company and moderators in future cases. In a sense, the Oversight Board represents another twist on the outsourcing theme, as Facebook seeks to shift responsibility for certain moderation decisions to an independent body, albeit one that it has created and financed.

From Porn to Violence

Outsourced Facebook moderation takes place at more than 20 sites worldwide, although the company won't confirm how many countries host these sites. Third-party vendors—companies such as Accenture, Competence Call Center, CPL Resources, Genpact, and Majorel—

run these operations. India, Ireland, and the Philippines host major hubs, each of which handles content automatically routed to it from all over the globe. Other Facebook sites, such as those in Kenya and Latvia, focus primarily on content from their respective regions. When they are called upon to review content in languages they don’t know, moderators use the company’s proprietary translation software. Facebook moderators collectively understand more than 50 languages, but the platform supports more than 100—a large gap, which the company says it is working to close. As noted earlier, Facebook declined our repeated requests to visit a moderation site and talk to current workers.

Christopher Gray left a temporary gig teaching English in China to become a Facebook moderator with CPL Resources in 2017 in Dublin, his adopted home city. His wife, who at the time worked for CPL in a non-moderator role, told him about the opening. CPL, which is based in Dublin and has operations across Europe, was expanding its work for Facebook. Gray, who is now 50, started with the relatively modest hourly wage of €12.98, or about $14. CPL also paid a small bonus for working nights and a daily travel allowance. An enthusiastic Facebook user, Gray was impressed by the company’s glass office building in Dublin’s bustling Docklands section. “It has a large atrium and lots of light, artwork, green plants, free food, coffee and tea,” he says in an interview. And he began with a sense of mission. “You have this feeling that you’re there to do good, protect the users,” he says. “We’re making a contribution.”

Facebook says in a company statement that moderators receive “extensive training” that includes “on-boarding, hands-on practice, and ongoing support and training.” Gray describes his training as only eight days of “pretty cursory” PowerPoint displays presented in rote fashion by a CPL staff member. On his first day of actual moderating, Gray went to work on a global pornography “queue.” He sat in front of a desktop monitor deciding whether one image after another violated Facebook’s prohibition of “adult nudity and sexual activity,” as defined in the Community Standards. Employees assigned to other queues assessed content that had been flagged for hate speech, graphic violence, terrorist propaganda, and so forth. The moderators’ options were to click “ignore,” “delete,” or “disturbing,” the last of which applied to items that didn’t violate the standards but might upset some users. Facebook covers disturbing content with a warning that users have to click through if they want to see it.

CPL “team leaders” set daily “game plans” for the number of pieces of content, or “tickets,” each moderator should process. The prescribed “average handling time” was 30 to 60 seconds per ticket, Gray says. That translated to 600 to 800 pieces of content over the course of a typical eight-hour shift, during which he would take time out for short breaks and a meal. When the queue got backed up, he says he would “blast through” 1,000 items in a shift. Facebook says that the company’s third-party contractors don’t enforce quotas and instead encourage reviewers to take as much time as they need to evaluate content.

Over the months, Gray’s circumstances changed. He and his working group were moved to a drab CPL-run building in Dublin, and he was switched from porn to a “high priority” queue containing a mixture of deeply troubling material requiring immediate attention. He’d shrugged off the pornography but now was confronted by imagery he couldn’t get out of his head, even during his off hours. He recalls one video in which men in black balaclavas used machine guns to mow down a group of captives in orange jumpsuits at point-blank range. In another, a woman wearing an abaya was stoned to death. In a third, an alleged sex offender in Russia was whipped to death. Animal torture was common, including a video showing dogs being cooked alive in China. He marked all of this content for deletion, but watching it unfold on his screen took a toll.

> **“One moderator recalls videos showing men in black balaclavas using machine guns to mow down captives in orange jumpsuits, a woman wearing an abaya being stoned to death, and an alleged sex offender in Russia getting whipped to death. Animal torture was common, including a video showing dogs being cooked alive in China.”**

Gray’s experience wasn’t unusual. In his first week working for CPL at about the same time, Sean Burke remembers watching a Facebook video of a man being beaten to death with a wooden board with nails sticking out of it. The next day, he encountered a bestiality video for the first time, and soon thereafter he started seeing child pornography. “They never actually teach you to process what you’re seeing,” says Burke, now 31. “It’s not normal seeing people getting their heads cut off or children being raped.” Some of the worst images were uploaded hundreds or even thousands of times, coming back over and over to haunt moderators like Burke and Gray, who often had to remove multiple copies of the same content.

As part of his supervisory duties, Guy Rosen, Facebook’s vice president for integrity, has immersed himself for hours at a time in raw content from moderation queues. Without reference to any specific reviewers, he acknowledges in an interview that being a full-time moderator would be arduous. “It’s a hard job,” he says. “It’s a really hard job.”

> “Accenture, which does content moderation for social media companies, had its employees sign a form that said, ‘It is possible that reviewing such content may impact my mental health, and could even lead to post-traumatic stress disorder (PTSD).’”

PTSD Symptoms

Gray and Burke, who didn’t know each other at the time they worked for CPL, say they gradually began suffering psychological symptoms they now associate with their social media employment. They experienced insomnia and nightmares, unwanted memories of troubling images, anxiety, depression, and emotional detachment. CPL provided a certain amount of health insurance, but the policy specifically excluded mental health coverage, according to a personal injury lawsuit Gray filed in December 2019 seeking damages from the vendor and Facebook. In the Irish High Court suit, Gray cited a physician’s diagnosis that he suffers from a form of post-traumatic stress disorder (PTSD). Burke, who says he also has a PTSD diagnosis from a physician, is considering taking similar legal action.

Gray worked for CPL for nine months; Burke, for about a year. They both were let go for allegedly failing to meet accuracy standards in their reviewing work—a topic that we’ll return to below.

Facebook has denied liability in the Irish case, and a spokesperson declines to comment. Offered multiple opportunities to discuss CPL’s work for Facebook and the Gray lawsuit, a CPL spokesperson says via email: “We do not comment on specific client engagements” and refuses to elaborate.

A similar lawsuit was filed as a class action against Facebook in state court in San Mateo County, California, in 2018. A group of former reviewers alleged that “as a result of constant and unmitigated exposure to highly toxic and extremely disturbing images,” they had suffered “significant psychological trauma and/or post-traumatic stress disorder.” Without admitting to any wrongdoing, Facebook agreed in May 2020 to settle the San Mateo suit in an out-of-court deal that could distribute tens of millions of dollars to more than 10,000 current and former moderators in the U.S.

All members of the class will be eligible for a $1,000 payment designed to cover medical screening for mental health problems. Beyond that basic amount, plaintiffs may qualify for up to $50,000 if they can produce evidence of serious harm, such as a doctor’s diagnosis of PTSD. Facebook agreed to provide a total of $52 million, out of which the plaintiffs’ attorneys’ fees also will be paid. In addition, the company agreed to ensure that its third-party vendors will provide more access to mental health coaching from licensed professionals. As of this writing, the settlement awaited approval by the judge supervising the case.[18]

Facebook put a notably positive spin on the out-of-court resolution in California. “We are grateful to the people who do this important work to make Facebook a safe environment for everyone,” the company said in a prepared statement. “We’re committed to providing them additional support through this settlement and in the future.”

Facebook’s implicit acknowledgment in the California settlement that content reviewers deserve better psychological care followed a different sort of tacit concession by Accenture, one of its third-party vendors. In December 2019 and January 2020, Accenture confirmed the potential health risks facing moderators when it told employees at Facebook and YouTube sites in Europe and the U.S. to sign a two-page form attesting to the dangers. “I understand the content I will be reviewing may be disturbing,” the form stated. “It is possible,” the form continued, “that reviewing such content may impact my mental health, and could even lead to post-traumatic stress disorder (PTSD).” Moderators, the form added, should take full advantage of workplace “wellness” programs. But the “wellness coach” provided by Accenture “is not a medical doctor and cannot diagnose or treat mental health disorders,” the company stated. Accenture seemed to suggest that some employees should consider just quitting. “No job is worth sacrificing my mental or emotional health,” the company’s form said.[19]

Facebook has nothing to say about the Accenture form. "I don't think I can comment on what they're doing," says Arun Chandra, Facebook's vice president for scaled operations. A YouTube spokesman similarly declines to comment, saying Accenture is best suited to respond.

Asked about the form, Accenture responds in a prepared statement: "We regularly update the information we give our people to ensure that they have a clear understanding of the work they do—and of the industry-leading wellness program and comprehensive support services we provide, which include proactive and on-demand coaching backed by a strong employee assistance program." There were "no consequences for not signing the updated document," the statement adds. Once known as Andersen Consulting, Accenture is based in Dublin and has more than 500,000 employees worldwide. "The well-being of our people is a top priority," the company says.

A 2019 study by scientists employed by Google provided additional confirmation of the mental health risks of doing content moderation. The authors noted "an increasing awareness and recognition that beyond mere unpleasantness, long-term or extensive viewing of such disturbing content can incur significant health consequences for those engaged in such tasks."[20]

Inadequate Care

At this point, it bears emphasizing that this report is *not* arguing that outsourcing in and of itself causes potentially serious psychological harm. This harm is intrinsic to the activity and could well afflict moderators working as full-time employees of Facebook, YouTube, or Twitter. The argument, instead, is that given the nature of the work in question, outsourced moderation operations tend to lack the medical care or counseling employees need to remain healthy.

As Accenture mentioned in its acknowledgment form, third-party vendors typically provide "wellness" programs, not true psychological counseling, let alone care from a medical doctor. Wellness coaches generally don't grapple with psychological harm. CPL "offered yoga and breathing exercises and finger painting, nothing that dealt directly with what we were seeing every shift," Gray recounts. Over the nine months he worked for CPL, he says he had three one-on-one sessions with a counselor, who repeatedly asked him about his "strategy" for dealing with the alarming material on his screen. "I didn't know my strategy," Gray says. "That's why I went to counseling. I was at a loss."

Burke says he made it to only one group wellness session where the counselor organized a "bonding" exercise in which employees were supposed to say one good thing about each other. "There wasn't any real focus on what we were moderating," Burke says. It was almost as if the counseling had been designed to avoid the real problem at hand, he adds. "They give you motivational speeches—take a deep breath, jump around, and loosen up—it was not psychological help," says Valera Zaicev, another moderator who worked for CPL in Dublin from 2016 through late 2018. Zaicev says he, too, is considering filing suit against Facebook and CPL.

Moderators elsewhere agree with these assessments. "The counseling was a joke," says Clifford Jeudy. He earned about $16 an hour working from 2017 through 2019 at a Facebook moderation site in Tampa run by Cognizant, a 290,000-employee company based in New Jersey. Among Jeudy's more grimly memorable experiences was having to watch the March 2019 massacre carried out by an avowed white supremacist who killed 51 worshipers at two New Zealand mosques. The killer live-streamed his attacks, and numerous other users re-uploaded versions of the video more than 1.5 million times on Facebook. Jeudy viewed and removed it multiple times. "There was no counselor on-site after I had to watch that," he says. "Often, the counselor wasn't even there, and when there was someone around, they didn't seem interested in helping you figure out how to cope with what you were seeing." Jeudy, now 47, says he was diagnosed with anxiety disorder and a form of PTSD in July 2019. Subsequently, he suffered a stroke, which he also attributes to stress from moderating.

> **"'They give you motivational speeches—take a deep breath, jump around, and loosen up—it was not psychological help,' one former content moderator says when describing counseling available through on-site wellness programs."**

A Cognizant spokesperson says via email that counselors were on-site in the Tampa facility "12 hours a day, seven days a week. Mental health professionals also provided daily check-ins and were accessible to all staff." An employee assistance program "provided access to licensed professionals in the community for employees and family members," the spokesperson adds. Cognizant also offered "a robust wellness program" that included yoga, mindfulness, and pet therapy, according to the spokesperson. Cognizant and Facebook are named as defendants in a lawsuit filed on behalf of a group of moderators that is pending in federal court in Tampa. Asked about the legal action, the Cognizant spokesperson says: "We do not comment on pending litigation and look forward to defending ourselves in the appropriate forum." Facebook also denies any wrongdoing in the Tampa case.

“At a Facebook moderation site in Tampa, the quality-audit process led to loud arguments, physical confrontations, and occasionally even a fist fight, says Debrynna Garrett, a former moderator who worked there until March 2020. ‘It is incredibly distracting to have that going on day after day,’ she says.”

Chandra, Facebook’s vice president for outsourcing, emphasizes that the company is working with its vendors to ensure that moderators receive necessary care. In 2019, the company announced new standards that include a requirement that vendors provide access to on-site counseling during all hours of operation and a 24-hour help hotline. Reviewers are supposed to get breaks whenever they need to step away from what they’re watching, and vendors are expected to set up special “resiliency areas” separate from the “production floor.” Facebook last year began commissioning independent audits of how vendors fulfill their obligations, but Chandra declines to discuss any initial findings. The company also is hosting periodic “partner summits” in Menlo Park, during which Facebook reinforces to vendor representatives the expectations it has for moderator support. “We’ve established a very high bar,” Chandra says.

But the bar for at least some full-time Facebook employees is higher. Workers at the Menlo Park headquarters have access to free on-site counseling at an in-house facility offering “individual therapy, psychiatry, groups, and classes,” according to the company website. All full-time Facebook employees have coverage for mental health therapy through the company’s several medical insurance programs. The company also offers 24/7 telephone counseling support for “in-the-moment assistance.”

Complementing moderator health benefits, Maxime Prades, a director of product management at Facebook, oversees the improvement of optional software features intended to lessen the impact of having to watch large amounts of disturbing content. These include the ability to blur alarming images, shift imagery from color to black and white, block faces, and mute sound. “We give them just what they need to make a decision,” Prades says in an interview.

In yet another move intended to improve the outsourced moderators’ experience, Facebook says that by the middle of this year, it will guarantee that in the U.S., these workers are paid higher minimum wages. In the San Francisco Bay area, New York, and Washington, D.C., the new floor will be $22 an hour, while in Seattle it will be $20, and in other U.S. metro areas, $18. These pay rates are well above legally mandated minimum wages.

The company says it is also considering whether to underwrite pay improvements overseas, where in some places, moderator compensation is markedly lower. *Reuters* reported in August 2019 that in Hyderabad, India, some outsourced Facebook moderators were paid starting annual salaries of 250,000 rupees, which works out to about $3,300 a year.[21]

Unsettled Workplaces

A separate problem that stems from the outsourcing of content moderation is the unruly work environment in which at least some moderators find themselves. “It interferes with the quality of the work almost every day,” says Clifford Jeudy, who worked for Cognizant in Tampa through 2019. “You’ve got to assume it would be different if Facebook was in charge.”

Former CPL employee Valera Zaicev, now 33, got a glimpse of what it would be like to be employed directly by Facebook in Dublin. During his first year, he worked in the glass Facebook building in the Docklands section. Fluent in Russian and English, he was assigned to a queue with potentially problematic material from the former Soviet republics. The content, which included apparent threats and hate speech, required painstaking review in multiple languages, and, as a result, he was expected to complete only 60 to 80 tickets a day. That left plenty of time during his eight-hour shift to visit the Facebook library, game room, or gym. More important, he says, he had the opportunity to interact on a daily basis with full-time Facebook employees who were expert in his area. “They were able to guide us and teach us,” he says. “It was very organized and civilized.”

Much of that changed, he says, when he was moved to a CPL-run building about two miles away. His work target jumped more than fourfold, to 350 tickets a day. "With that kind of increased productivity demand comes a decline in quality," he says. Typically, he adds, moderators under this kind of pressure leave up material that perhaps should come down. The content can go viral before anyone has a chance to give it a second look.

Exacerbating the situation, he says, was a lack of access to Facebook personnel who knew what they were talking about. "CPL's team leaders and trainers had no experience," he explains. "They really didn't know more than we did; sometimes, less." Specifically, the CPL supervisors "didn't know the slurs, the political tensions, the dangerous organizations, or the terrorist organizations" characteristic of the former Soviet republics.

CPL quality auditors, known as QAs, stepped up the intensity of their reviews, leading to daily squabbles over whether Zaicev's decisions were correct. "You get the feeling you are a second-class citizen," he says. "They don't take you seriously. As a result, you don't do your best work."

Facebook generally sets a goal for moderation sites of 98% "accuracy" rates, as determined by the QA review process. Individual moderators are also judged on their accuracy rating. Figuring into that rating is not only a moderator's bottom-line decision to remove or allow a piece of content, but also the reason for the decision, chosen from a long drop-down menu. Select the "wrong" reason, and the QA marks down an error. The Dublin moderators Gray and Burke say they were told they were let go because their accuracy rates slipped to the mid-90s.

Internal auditing to encourage diligence and accuracy is, of course, a good idea. But QA scores are much less reliable than they may seem, according to moderators. That's because the QAs' calls often amount to highly subjective second-guessing. Moderators frequently dispute the QAs' judgments in a ritual known as "getting your points back." For example, Gray once encountered an image of a baby with its eyes closed, its arms hanging limply at its sides, and an adult foot pressing down on its chest. Interpreting this grotesque scene as depicting a violent death, he deleted the content. The QA disagreed, leading to an extended back-and-forth over whether there was sufficient evidence that the baby was, in fact, dead. The QA wouldn't back down, and Gray's decision was marked as incorrect. The disturbing image was restored.

In Tampa, the QA-dispute process led to loud arguments, physical confrontations, and occasionally even a fist fight, says Debrynna Garrett, a former police officer who worked until March 2020 for Cognizant as a Facebook moderator. "It is incredibly distracting to have that going on day after day," she says. "The whole process was the blind leading the blind," adds her ex-colleague Jeudy. "The accuracy scores were never accurate because we got back so many points during the dispute process." He estimates he had two-thirds of his "errors" overturned. Garrett is a plaintiff in the federal court case in Tampa against Facebook and Cognizant.

Some moderators react to their difficult office environment by trying to divert themselves during working hours. In February 2019, *Bloomberg News* interviewed several moderators anonymously about conditions at a site in Austin, Texas, operated by Accenture on behalf of Facebook. Many moderators, the news service reported, "attempt to pay as little attention to their work as possible, either by listening to music or streaming movies as they work." Alcohol and marijuana use by moderators is common both on and off the Austin premises, *Bloomberg* added.[22] Asked about this account, Accenture doesn't address it in its emailed response.

In a three-part series posted over the course of 2019, the online tech magazine *The Verge* described harrowing conditions in several U.S. moderation sites, including Cognizant centers in Tampa and Phoenix. One of many examples: After colleagues physically threatened him during multiple disputes, a QA in Phoenix began bringing a gun to work for self-protection, *The Verge* reported.[23]

The Cognizant spokesperson says that the company "strives to create a safe and empowering workplace for its employees around the world. Like any large employer, Cognizant routinely and professionally responds to and addresses general workplace and personnel issues in its facilities." The spokesperson adds that "Cognizant takes threats of workplace violence seriously and investigates all complaints thoroughly. If we find an issue, we take appropriate action, including termination."

Cognizant announced in October 2019 that it would exit the content moderation business because the activity no longer fit with the company's "strategic vision." Facebook has said that to compensate, it will expand moderation operations in Texas run by another company.[24]

Facebook's Response

Facebook says its outsourced operations typically run smoothly. Chandra, the vice president overseeing the area, emphasizes that the social media company doesn't directly supervise its outsourced content moderators. "We don't drive the day-to-day activities of the reviewers," he says. "Because ultimately the day-to-day management is done by the partners themselves"—companies like CPL, Accenture, and formerly Cognizant.

"When you run an operation of this size, will there be situations that come up that are not optimal? Yes," Chandra adds. When Facebook hears about a problem site, he says, "we investigate it fully and thoroughly." In general, the executive adds, negative reports about outsourced operations aren't consistent with what he's observed during visits to the sites since he joined Facebook in late 2018. "Candidly, many of them are better than Facebook offices," he says. Moreover, "the feedback I hear from our partners is that we are leading the industry in terms of the work that we are driving."

On these site visits, he says he makes a point of taking aside content moderators so they can speak out-of-earshot of their bosses. What he hears is pride, not complaint: "There are so many people who have said that by doing this kind of work, they are helping their kids stay safe or their community stay safe because [harmful] stuff has been taken off the platform." Chandra says this sense of mission helps explain the low attrition rate among moderators. Facebook, however, declines to specify this rate, and anecdotal evidence suggests that moderators rarely last more than a year or two in the job.

Asked why, if content moderation is a vital corporate function, Facebook outsources it, Chandra says, "That's a question I often get." Big corporations frequently outsource functions that aren't in their area of core expertise, he says. "Today, if you go to any large company, any Fortune 500 company, you'll find that their finance back office is all outsourced."

But if financial back-office activities were as traumatic for some workers as content moderation, and took place in as disruptive an atmosphere, other Fortune 500 companies would be obliged to rethink their outsourcing strategies. Furthermore, content moderation seems so important to running Facebook that it ought to be regarded as falling within the company's core activities, not as an ancillary chore to be handled by contractors.

Facebook sees advantages in outsourcing, Chandra continues. By contracting to have moderators hired around the globe, Facebook is able to meet its need for people who are fluent in dozens of languages. Global reach also allows the Facebook moderation operation to function in nearly every time zone, 24 hours a day, seven days a week. Finally, the arrangement allows Facebook to shift resources quickly in response to changing needs. Referring to recent events, Chandra says that when political turmoil in Hong Kong and Iran spurred increased reports of hate speech and misinformation, Facebook's third-party vendors were able to assign personnel to focus on these hot spots.

It makes sense for Facebook to have global coverage in every sense of the term. But this coverage would work better if the company hired full-time Facebook employees as moderators and positioned them in Facebook offices around the world. Doing so would require Facebook to extend itself in terms of various human resources functions—recruitment, training, and so forth—and it would require a significant financial investment, reducing the company's profits. But those profits, $18.5 billion on $70.7 billion in revenue in 2019, are robust enough to allow the company to achieve better, more humane content moderation.

> **"**
>
> **'There are so many people who have said that by doing this kind of work, they are helping their kids stay safe or their community stay safe because [harmful] stuff has been taken off the platform.' — Arun Chandra, Facebook's vice president for scaled operations**
>
> **"**

4. Content Moderation and Volatile Countries

Pursuing a business strategy of aggressive global growth, Facebook has created the difficulty for itself of moderating the prodigious output of its 2.5 billion users worldwide. These users express themselves in more than 100 languages and in all kinds of cultural contexts, making effective content moderation exceedingly difficult. Twitter and YouTube face similar challenges. To handle the crucial responsibility of content moderation, Facebook and its rivals have chosen to largely outsource the human part of the function, leading to the problems examined earlier in this report. These are not, however, the only problems related to the outsourcing of moderation.

Other harms—in some cases, lethal in nature—have spread as a result of Facebook's failure to ensure adequate moderation for non-Western countries that are in varying degrees of turmoil. In these countries, the platform, and/or its affiliated messaging service WhatsApp, have become important means of communication and advocacy but also vehicles to incite hatred and in some instances, violence. Myanmar is the most notorious example of this phenomenon; others include Cameroon, the Central African Republic, Ethiopia, Nigeria, Sri Lanka, and India. In some of these places, Facebook at times has, in effect, outsourced monitoring of the platform to local users and civil society organizations, relying too heavily on activists as a substitute for paid, full-time content reviewers and on-the-ground staff members.

"We were slow to identify these issues," says Guy Rosen, Facebook's vice president for integrity. "Unfortunately, it took a lot of criticism to get us to realize, 'This is something big. We need to pivot.'"

> "'We were slow to identify these issues' about at-risk countries, says Guy Rosen, Facebook's vice president for integrity. 'Unfortunately, it took a lot of criticism to get us to realize, "This is something big. We need to pivot."'"

The outside criticism, especially on Myanmar, launched a three-year corporate "journey," Rosen adds. He and his fellow executives contend the experience has made Facebook in 2020 far more vigilant and better prepared to respond to crises in at-risk countries. In 2018, the company formed a Strategic Response team, based in Menlo Park but prepared to swoop into such countries at the first hint of trouble. Reporting directly to Chief Operating Officer Sheryl Sandberg, the number two executive at the company, the team advises on such issues as where more moderators may be needed and consults with engineers on how to adjust technology to minimize rumors and misinformation that can lead to violence. The Strategic Response team, whose size Facebook declines to reveal, helped launch a feature for Sri Lanka that restricts users to sharing only posts from their Facebook friends. By "adding more friction," team member Sarah Oh says in an interview, the platform prevents some incendiary content from going viral.

Facebook also has adjusted its Community Standards to address risks in volatile countries. Under one recalibrated policy, content moderators now are supposed to remove "verified misinformation" and "unverifiable rumors" that may contribute to imminent physical harm offline. In 2019, Facebook hired its first human rights director, Miranda Sissons, a well-regarded veteran in the field who has worked in government and for civil society organizations and has experience in technology. In May 2020, Sissons oversaw the public disclosure of summaries of a series of "human rights impact assessments" of Facebook's role in Sri Lanka, Indonesia, and Cambodia.

Even with the company's changed consciousness and new hires, the question remains whether Facebook has expanded into more countries than it's prepared to safeguard from potential misuse of its platform. The goal in considering this question shouldn't be that Facebook curtail its reach by cutting off developing countries. These are places where the platform has become a vital organizing and communication tool for dissenters and civil society groups, as illustrated during the Arab Spring in the early 2010s and many times since then. Facebook also provides a virtual marketplace where aspiring entrepreneurs can promote small businesses. But wherever it operates, Facebook needs to buttress its services with adequate content moderation—preferably carried out by full-time Facebook employees who themselves receive the sort of support they deserve. As we explain in our first recommendation on page 24, these moderator-employees need not necessarily be located in Silicon Valley; often it will make more sense for them to be based in or near the countries whose content they are reviewing.

The following are five illustrative examples—not a comprehensive list—of countries where misuse of Facebook has been associated with ethnic or religious strife:

Myanmar

In March 2018, United Nations investigators declared that Facebook had played a "determining role" in the ethnic cleansing of Myanmar's minority Rohingya Muslims by the country's military and allied Buddhist groups. As many as 10,000 Rohingya were killed, and more than 700,000 fled as refugees to neighboring Bangladesh.

A majority-Buddhist country formerly known as Burma, Myanmar has a population of 54 million people. Twenty-three million use Facebook, representing more than 90 percent of those who are online. But for years, Facebook had all but ignored anti-Rohingya propaganda spread on its platform. During the ethnic cleansing, there were no more than a few Burmese-speaking moderators working on an outsourced basis from outside the country. Local civil society groups conveyed warnings to Facebook, which at first weren't heeded. In April 2018, Mark Zuckerberg testified before the U.S. Senate: "What's happening in Myanmar is a terrible tragedy, and we need to do more."[25]

Over the past two years, Facebook has done more. The company says it has seen to the hiring of more than 100 Burmese-language moderators. It has also invested in Burmese-language hate speech "classifiers," which are a type of AI-driven technology designed to detect offending expression proactively. To achieve this enhanced detection capability, the company had to convert Myanmar's unique system of language characters into the global standard known as unicode. Facebook has instituted similarly customized hate speech detection technology for other at-risk countries, as well.

Seeking to prevent incitement to violence, Facebook also has removed hundreds of pages, groups, and accounts, including many linked to the Myanmar military. And it has stepped up its interaction with local human rights activists and organizations, whose alerts to Facebook are now supposed to be prioritized by specially devised reporting technology.

But some local observers say that, given continuing ethnic and religious volatility in Myanmar, the company's progress hasn't gone far enough. Facebook remains "secretive" about where its Myanmar moderators are based and how many of them are devoted to the country's numerous minority languages, as compared to the volume of content in each of those languages, Aye Min Thant, a project manager at the tech-oriented NGO Phandeeyar, says in an interview. In addition, when Facebook removes content, it doesn't provide warning before deleting suspect pages, groups, and accounts, she says. This can interfere with pending investigations by civil society organizations and inhibit broader efforts to identify material that could provoke ethnic strife. The main complaint, though, is that Facebook doesn't have an office or locally based full-time staff in Myanmar. The company ought to be able to see over the horizon to detect the beginnings of any new forms of misuse of its platform, says Jes Kaliebe Petersen, who heads Phandeeyar. "That's very hard to do when they don't have a presence here."

Facebook responds that while the locations of some of its moderation sites are known—Dublin and Manila, for example—"we have refrained from sharing too many details about our sites and the people that work there due to security concerns."

India

Some Rohingya chased from their homes in Myanmar have ended up in India, where they've become the target of Hindu nationalist hatred. Once again, the antagonists, some of them affiliated with Prime Minister Narendra Modi's Bharatia Janata Party (BJP), have exploited Facebook in one component of a broader anti-Muslim movement in India. The majority-Hindu country is Facebook's largest market, with some 300 million users out of a total population of 1.4 billion.

In one video circulated widely in 2019, a group of men affiliated with the militant wing of the BJP brandished knives and burned the effigy of a child while screaming, "Rohingyas, go back!" in Hindi and

English. Other posts showed gruesome images of human body parts and falsely claimed that the Rohingya are cannibals. Facebook reportedly did not remove the burning effigy video on the theory that it was posted by groups claiming to be news organizations and wasn't directly linked to violence. The link may not have been direct, but in June 2019, dozens of Rohingya homes were burned in Jammu, where the video and others like it were shot. The cannibalism accusations were often removed because they violated Facebook's standards against graphic violence and hate speech, but the offending posts nevertheless kept reappearing, according to *The New York Times*.[26]

In the northeastern state of Assam, researchers for the advocacy group Avaaz identified a menacing social media campaign against Bengali Muslims, who were called "parasites," "rats," and "rapists." Avaaz said in a report published in October 2019 that it had flagged 213 of "the clearest examples of hate speech" directly to Facebook but that the company had removed only 96. The disturbing posts "were easily found by native Assamese speakers, and yet Facebook's own team had not previously detected any of them before being alerted to them by Avaaz," the group said in its report.[27]

The South Asian advocacy group Equality Labs published a separate report in June 2019 that described hundreds of memes and posts targeting Indian caste, religious, and LGBT minorities. Equality Labs brought this content to Facebook's attention over the course of a year but the platform failed to remove it. Religious-themed posts attacked Muslims and Indian Christians by depicting verbal and physical bullying, burning of Bibles, and exhortations to violence. A meme advocating domestic violence, without reference to religion, showed a pair of male hands gripping a baseball bat and accompanied by the text, "Did you know that this object, besides being an educational tool for wives, is also used in an American game called baseball?"[28]

In response to challenges such as these, Facebook is paying more attention to content moderation in India and other at-risk countries, spokeswoman Ruchika Budhraja says via email. "We've seen significant increases in the number of reviewers with language expertise" in recent years in India and such countries as Sri Lanka, Indonesia, and Ethiopia, which are discussed below. "But we don't typically provide country-by-country breakdowns," she adds. Myanmar is an exception because of the nature and degree of Facebook's involvement in the human rights crisis there and the extensive public attention it drew. Budhraja says that identifying the number of moderators tracking content in a given country provides a "wholly inaccurate reflection of the number of people and resources focused on any specific country." That's because many others in the company, in addition to moderators—including engineers, product and policy staff, and Strategic Response personnel—are involved, she says. And the ranks of these other categories have grown, too.

Still, the human rights consulting firm Article One Advisors argues for more disclosure related to content moderators. In a human rights review commissioned by Facebook and released in May 2020, one of Article One's recommendations was that the social media company should "publish data on content moderators, including the number of content moderators disaggregated by language expertise, country of origin, age, gender, and location." Facebook's resistance to providing even the basic number of moderators in particular countries obviously makes it more difficult to evaluate its efforts on this front.[29]

Sri Lanka

India's much smaller neighbor Sri Lanka, with a population of 21 million, about a third of whom are Facebook users, has seen both anti-Muslim and anti-non-Muslim violence fed by Facebook content.

In March 2018, Sinhalese Buddhists rioted and torched homes and businesses in the central district of Kandy, leaving three people dead and the country under a state of emergency. Among the Facebook posts that fanned ethnic and religious animosity were allegations of a nonexistent Muslim plot to wipe out the country's Buddhist majority and violent exhortations in the Sinhala language. The calls to violence included, "Kill all Muslims; don't even save an infant. They are dogs." Local Sri Lankan activists informed Facebook of the building antagonism, the *Times* reported, but received the baffling response that the vitriolic content didn't violate the company's standards.[30] In its recently released human rights impact assessment, Article One concluded that "the Facebook platform contributed to spreading rumors and hate speech, which may have led to 'offline' violence."[31]

A little more than a year after the anti-Muslim riots, religious violence encouraged by social media fulmination again shook Sri Lanka. A local Islamic leader who allegedly helped plan bombings coinciding with Easter in 2019 used Facebook to call for killing non-Muslims. The attacks on churches and hotels resulted in the death of more than 250 people and injury of hundreds more. Islamic preacher Zahran Hashim, who blew himself up, had said in one video posted seven months before the bombings: "Non-Muslims and people who don't accept Muslims can be killed along with women and children." Moderate Muslims had made concerted efforts to warn Facebook about Hashim's incendiary Tamil-language posts, according to *The Wall Street Journal*, and Facebook removed much of the content before the carnage. But about 10 posts, including the exhortation to kill non-Muslim women and children, remained visible on the platform through the time of the attacks and for days afterward.[32]

In its May 2020 human rights impact assessment, Article One noted that, at least in the past, technology has played a part in spreading hate speech on Facebook. The company's limited cultural and language expertise, the consulting firm said, "was potentially exacerbated by now-phased-out algorithms designed to drive engagement on the platform, regardless of the veracity or intention of the content."[33] Asked to respond, a Facebook spokesperson declined to comment.

Facebook, in its written response to the Article One assessment, noted that since 2018, it has hired policy and program managers to work full-time with local stakeholders in Sri Lanka. It has also "deployed proactive hate speech detection technology in Sinhala." And without providing precise numbers, the company said that "dozens" of additional Sinhala- and Tamil-speaking content moderators are now reviewing Sri Lankan content.[34]

Indonesia

Indonesia is the world's fourth-most-populous country, with more than 270 million people, nearly half of whom are Facebook users. The southeast Asian nation is also the world's most populous Muslim-majority country. In a separate assessment also released in May 2020, Article One described Indonesia as having a "mixed record on human rights." For example, the country has enacted legal provisions protecting free expression but also enforces harsh criminal anti-defamation and anti-blasphemy laws, especially against non-Muslims, based on their use of Facebook and other social media platforms.[35]

Another problem is the proliferation on Facebook, WhatsApp, and to a lesser extent, Instagram of misinformation and disinformation, or "hoaxes," as they're known in Indonesia. "Hoax factories," such as one called the Muslim Cyber Army, have targeted political figures and promoted disinformation in hopes of tilting the results in local and national elections, according to Article One. One Indonesian study cited by the consulting firm found that 92% of respondents said they had received hoaxes from platforms like Facebook, Twitter, Instagram, and YouTube.[36]

Facebook's response to these instances of platform misuse "was slow and, at times, insufficient—potentially exacerbating impacts," Article One determined. The consulting firm's analysis was limited to events occurring through December 2018. But the problems have continued, including during national elections in 2019, when a heavy flow of misinformation focused on "stoking ethnic and religious divides while depicting electoral bodies as corrupt," *Reuters* reported.[37]

For its part, Facebook says it has made numerous oversight improvements in Indonesia, most of which are similar to changes mentioned previously in this section. For instance, the company has increased the number of content moderators and added policy personnel. It has rolled out proactive hate speech detection technology in the Bahasa Indonesia language. And it has taken steps to limit the viral potential of forwarded messages on WhatsApp.

Ethiopia

A multi-ethnic society with a population of 114 million, some 6 million of whom use Facebook, Ethiopia illustrates both the promise and the peril of social media in developing countries.

Since Prime Minister Abiy Ahmed took office in 2018 and rolled back long-standing repressive policies, Facebook has served as an important tool for civil society advocates and journalists. But at the same time, the platform has become a venue for exacerbating the sometimes-bloody rivalries among the country's regional-ethnic groups. Abiy referred to the situation in Stockholm in December 2019 during his acceptance speech for the Nobel Peace Prize, which he won for resolving hostilities with neighboring Eritrea. "The evangelists of hate and division," he said, "are wreaking havoc in our society using social media."[38]

One incident Abiy surely had in mind had occurred two months earlier, when one of his political rivals, Jawar Mohammed, posted on his widely followed Facebook account that government authorities had launched a plot against his life by seeking to remove his officially provided bodyguards. His supporters from the Oromo ethnic group rallied to his defense, and this activity devolved into communal violence, as graphic images purporting to show the results of the protests circulated on social media. Ultimately, an estimated 86 people died in the unrest. To be fair, Facebook may not have had an immediate basis for blocking the original post, but once violence broke out, nimble content moderators could have intervened to remove follow-up content that encouraged physical conflict.[39]

In a separate episode in June 2019, an army general named Asamnew Tsige used a video he spread via Facebook to call on the Amhara people to take up weapons for fighting against rival groups, according to *Reuters*. This call to arms led to an abortive coup which took the lives of the Ethiopian army's chief of staff and the head of the northern state of Amhara. Asamnew was killed and more than 180 people were arrested in the government's suppression of the coup.[40]

Reacting to these and other incidents, Ethiopia's parliament in February 2020 passed a law intended to curb social media excesses. The measure imposes fines and prison terms for people whose posts are deemed to have stirred unrest or violence. But the government could also use the law to go after legitimate dissenters relying on social media to promote their message. While it certainly can't be blamed as the sole reason for the problematic legislation, the lack of vigorous content moderation in Ethiopia appears to have contributed to a social media environment in which such a measure would gain political support.[41]

Many Frauds, Not Enough Fact-Checkers

In December 2016, after the Russian election-interference debacle, Facebook agreed to a partnership under which five outside fact-checking organizations would review content on the platform that had been flagged as possibly false. This partnership has since grown to include more than 60 paid fact-checking groups worldwide, which work in more than 50 languages and now review Instagram posts, as well. But the amount of potentially untrue content is vast, and fact-checkers say many dubious items go unevaluated. "The scale is so mind-boggling," Karen Rebelo, the deputy editor of Boom Live, a 10-person operation in Mumbai, India, says in an interview. "There is no way to keep up, even though we try very hard."

To address the enormous scale and catch at least some of the falsehoods that now slip through the cracks, Facebook needs to expand its international fact-checking program much further, ensuring that far more reporters are looking at a greater amount of suspect content. YouTube and Twitter, which do fact-checking on a much more limited basis, should launch similar programs of their own. (For more on this recommendation, please see page 26.)

The social media companies will never wipe their sites clean of all falsehoods. But the impossibility of achieving perfection with fact-checking shouldn't stand in the way of incremental improvement.

Fact-checking is a close cousin to content moderation. Rather than assess whether content should be removed because it violates Facebook's Community Standards—on bullying, say, or hate speech—fact-checking evaluates whether material is untrue. If fact-checkers successfully debunk an item, Facebook will label it as false and rank it lower in News Feed so fewer users will see and share it. Facebook also appends to the false content any article that the fact-checkers write providing more context on the topic. Facebook punishes pages and websites that repeatedly share false news by limiting their distribution and blocking their ability to monetize or advertise on the platform.[1]

Fact-checking often comes into play in the political realm, but it doesn't always have the desired effect. In a notorious example from May 2019, a conservative activist manipulated a video to make it appear that House Speaker Nancy Pelosi drunkenly slurred her words during a speech in Washington, D.C. Several fact-checking organizations branded the doctored video as false, and Facebook down-ranked it. In this instance, however, the video went viral anyway, garnering millions of views. Facebook should have limited circulation of the fraudulent content by removing it altogether, but that's not the company's policy. Twitter likewise did not remove the Pelosi video, although YouTube did take it down for being deceptive.[2]

Elsewhere, we've argued that the social media companies ought to remove demonstrably false content, prioritizing untruths related to democratic institutions, such as elections. We've also suggested that the platforms retain clearly marked copies of removed material in a searchable archive where it can be studied but not shared.[3]

Facebook fact-checkers range from large international journalism organizations like Agence France-Presse and Reuters to small groups that specialize in the activity, such as PolitiFact, a nonprofit operated by the Poynter Institute in St. Petersburg, Florida. Typically, each fact-checking organization assigns only a handful of reporters to investigate the veracity of Facebook content. PolitiFact's executive director, Aaron Sharockman, designates four of his group's 12 reporters to handle Facebook matters.

The social media company feeds posts to fact-checking organizations via a proprietary software "dashboard." Some of these candidates for potential debunking have been reported by Facebook users as suspect. Others have been identified by AI-driven screening technology, based on "signals" such as having been shared by a page that has shared misinformation in the past or attracting comments suggesting disbelief. PolitiFact typically assesses 2,000 to 3,000 Facebook posts a day, selecting only a few to investigate. Reporters ordinarily spend one or two days looking into and writing a fact-check, completing about 20 to 25 Facebook checks a week. Like some of its peers, PolitiFact also does fact-checks for other paying clients and pursues some leads that come in from readers. "Facebook has no say in our rating; nor do they have any say in what we pick" to evaluate, Sharockman adds.

Facebook pays by the piece, with a monthly cap on the number of paid fact-checks a given organization may do. Facebook requires fact-checking groups to sign nondisclosure agreements making the payment terms confidential, but some information has leaked out. *The Wall Street Journal* reported that Lead Stories, a Los Angeles-based organization with 10 full-time and six part-time staff members, earned $359,000 under its 2019 Facebook contract and expects to take in "a multiple" of that in the 2020 election year.[4]

Fact-checkers generally express enthusiasm for their mission but also a sense that too many falsehoods slip past them. "Frankly speaking," says Rakesh Dubbudu, founder of the Indian organization Factly, "I don't think fact-checking is sustainable by itself" to address the volume of questionable content.

[1] "Fact-Checking on Facebook," Facebook, undated, https://www.facebook.com/help/publisher/182222309230722.

[2] Sarah Mervosh, "Distorted Videos of Nancy Pelosi Spread on Facebook and Twitter, Helped by Trump," *The New York Times*, May 24, 2019, https://www.nytimes.com/2019/05/24/us/politics/pelosi-doctored-video.html.

[3] Paul M. Barrett, "Disinformation and the 2020 Election: How the Social Media Companies Should Prepare," New York University Stern Center for Business and Human Rights, September 2019, https://bhr.stern.nyu.edu/tech-disinfo-and-2020-election.

[4] Jeff Horwitz, "Facebook's Fact-Checkers Fight Surge in Fake Coronavirus Claims," *The Wall Street Journal*, March 30, 2020, https://www.wsj.com/articles/facebooks-fact-checkers-fight-surge-in-fake-coronavirus-claims-11585580400.

Recommendations

We intend this report not only as a critique of social media industry practices, but also as a basis for advocating constructive change. What follows are a series of proposals for how Facebook—and by extension, Twitter and YouTube—can improve the crucial functions of content moderation and fact-checking.

1 End outsourcing of content moderators and raise their station in the workplace.

Content moderation will improve if the people doing it are treated as full-fledged employees of the platforms for which they now work indirectly. Each of these companies should gradually bring on board, with suitable salaries and benefits, a staff of content moderators drawn from the existing corps of outsourced workers and others who want to compete for these improved jobs. This will require a significant investment by Facebook and the other social media companies. The added outlays will reflect the true cost of doing business responsibly as a global social media platform.

We're not suggesting that Facebook, YouTube, and Twitter need to employ all of their moderators in the U.S., let alone at their headquarters in northern California. We understand the advantage in having reviewers based around the world who bring to the job varied language skills and awareness of local cultures. This arrangement should be preserved and built upon—but with far-flung moderators looking directly to Silicon Valley for their supervision, compensation, and overall office well-being.

2 Double the number of moderators to improve the quality of content review.

In the past three years, Facebook has more than tripled the ranks of its reviewers. And one gets the sense that some senior executives worry about diminishing returns. "We need to be running a tight ship," says Guy Rosen, the vice president for integrity. "We need to make sure that we don't get lazy and just send more and more content to more reviewers, but that we're actually really thinking about how we're best using the skill and the time and the mental health of this person that's sitting and reviewing content for us."

But the reality is that Facebook and the other companies have more and more content and need to send it to more moderators. If, as a starting point, Facebook doubled its current number, to 30,000, it could give the expanded review workforce more time to consider difficult content decisions while still making these calls in a prompt manner. A larger moderator corps would also allow the platforms to rotate assignments more frequently so that reviewers exposed to the most troubling content could switch to queues with less disturbing material. A more sizable contingent would permit them to take moderators out of rotation on a daily basis for periods of more sustained, proactive counseling designed to determine early on whether they are in danger of developing PTSD or other psychological side effects from prolonged viewing of harmful content.

3 Hire a content overseer.

Facebook—and Twitter and YouTube—each should appoint a senior official to oversee the policies and execution of content moderation. The same person should also supervise fact-checking. As we have argued in past reports, responsibility for content decisions now tends to be scattered among disparate teams within the social media companies. Centralization would streamline key processes and underscore their importance internally and externally.

We suggest that to give the post heft, the new official report directly to the CEO or COO. One potential criterion for candidates for this job is experience in the news business, either as a top editor or executive. Someone from the serious side of journalism ought to have the right combination of acuity and commonsense about what sort of material belongs on a major social media platform.

4 Further expand moderation in at-risk countries in Asia, Africa, and elsewhere.

Facebook has arranged for additional outsourced moderators to pay attention to countries like Myanmar. Indonesia, and Ethiopia. This is a step in the right direction, and the expansion should continue until these countries have adequate coverage from moderators who know local languages and cultures—and function as full-time Facebook employees. Increased moderation needs to be accompanied by the presence of a country director and policy staff members in each country where Facebook operates.

Responsible global companies have people on the ground where they do business. A social media platform should be no different. Facebook, YouTube, and Twitter should have offices in every country where users can access their sites.

5 Provide all moderators with top-quality, on-site medical care.

People who hold what *The Wall Street Journal* once called "the worst job in technology"[43] deserve the best medical care to diagnose and treat psychiatric ailments brought on, or exacerbated, by extensive exposure to alarming content. This care should augment the wellness counseling and activities currently on offer by third-party vendors, and it should include access to well-qualified psychiatrists who have the authority to prescribe medication if that's indicated.

Because the invisible wounds caused by content moderation do not necessarily heal on the day that a moderator leaves the job, Facebook and the other social media platforms should provide generous, inexpensive health plans that continue for a period of years or until a worker obtains coverage via another employer. Casey Newton of *The Verge* argues sensibly that "tech companies need to treat these workers like the U.S. government treats veterans," offering them free, or heavily subsidized, mental health care for some extended period after they leave the job.[44]

6 Sponsor research into the health risks of content moderation.

The social media companies apparently don't know the precise risks their moderators take. Neither does anyone else. As third-party vendor Accenture has acknowledged, PTSD is one potential hazard, but how often does the affliction occur? Is there some weekly, monthly, or lifetime limit to the number of hours a moderator can safely watch harmful content? Perhaps such limits should define the length of a moderator's tour of duty and mark the point at which another assignment becomes available.

To answer these questions, Facebook and other social media companies that use content moderation ought to pool their considerable resources and underwrite wide-ranging, high-quality academic research.

(continued on p. 26)

Recommendations (continued)

7 Explore narrowly tailored government regulation.

In general, we oppose the sort of content regulation enacted in Germany, Singapore, and other countries which requires platforms to remove harmful content under threat of fines or other sanctions. Such laws constitute a form of government censorship and in the U.S. would be regarded as violating the First Amendment of the Constitution. For different reasons, we're also wary of politically charged proposals to rescind Section 230 of the Communications Decency Act, the law that shields platforms from most kinds of civil lawsuits over user-generated content.

In contrast, an idea that comes from Facebook itself is worth exploring. In an op-ed piece and a white paper, the company has proposed increased attention to the "prevalence" of harmful content. Facebook defines this term as the frequency with which deleterious material is viewed, even after moderators have tried to weed it out. Facebook proposes that a "third-party body"—possibly one created with government participation—could establish prevalence standards for comparable platforms. If a company's prevalence metric rose above a preset threshold, according to Facebook, it "might be subject to greater oversight, specific improvement plans, or—in the case of repeated systematic failures—fines." Structured properly, this approach would avoid government censorship of particular pieces of content and instead would create incentives for companies to curtail categories of material they themselves have identified as detrimental. Facebook has begun to publish prevalence numbers for certain types of malign content (see the chart on page 10). While companies might be tempted to game this system by minimizing their prevalence counts, the concept merits study and debate.[45]

8 Significantly expand fact-checking to debunk mis- and disinformation.

Disproving coronavirus conspiracy theories and other hoaxes, as well as politically motivated disinformation, is a noble pursuit but one that's presently done on too small a scale. Facebook has the most ambitious fact-checking program and deserves credit for that. But all of the social media platforms need to do more in this area—both to correct the record about particular demonstrably false posts and to send a broader message about the difference between fact and fabrication. Respect for this distinction is important to the health of democracy and especially to informed voting. The brutal contraction of the traditional journalism business in recent years means there are plenty of experienced former reporters who could be hired as fact-checkers.

Unlike content moderation, which would benefit from being brought in-house, fact-checking is best left in the hands of independent journalism organizations and specialty websites. Fact-checking gains credibility from the reality and perception of autonomy. Journalists, which is what fact-checkers essentially are, do their best investigating when they are not beholden to outside influences or institutions. If Facebook made it known that it wants to hire more outside fact-checkers, new and existing providers would step forward. YouTube and Twitter, which do far more limited versions of fact-checking, should emulate Facebook. Expansion of fact-checking would make the existing certification role that the International Fact-Checking Network plays all the more important. The nonprofit network promotes transparency and nonpartisanship in organizations it approves. Both values should remain hallmarks of social media fact-checking.[46]

Endnotes

[1] Tarleton Gillespie, Custodians of the Internet, Yale University Press, 2018, https://yalebooks.yale.edu/book/9780300173130/custodians-internet.

[2] Rob Price, "Facebook Moderators Are in Revolt Over Inhumane Working Conditions that They Say Erode Their 'Sense of Humanity,'" *Business Insider*, February 15, 2019, https://www.businessinsider.com/facebook-moderators-complain-big-brother-rules-accenture-austin-2019-2.

[3] Vijaya Gadde and Matt Derella, "An Update on Our Continuity Strategy During COVID-19," Twitter, March 16, 2020, https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html.

[4] Joshua Brustein, "Facebook Grappling with Employee Anger Over Moderator Conditions," *Bloomberg*, February 25, 2019, https://www.bloomberg.com/news/articles/2019-02-25/facebook-grappling-with-employee-anger-over-moderator-conditions.

[5] Sarah T. Roberts, Behind the Screen: Content Moderation in the Shadow of Social Media, Yale University Press, 2019, https://yalebooks.yale.edu/book/9780300235883/behind-screen.

[6] Jennifer Grygiel, "Are Social Media Companies Motivated to Be Good Corporate Citizens?" *Telecommunications Policy*, June 2019, https://www.sciencedirect.com/science/article/abs/pii/S0308596118304178?via%3Dihub.

[7] Mark Zuckerberg, "Building Global Community," Facebook, February 16, 2017, https://www.facebook.com/notes/mark-zuckerberg/building-global-community/10154544292806634/. I derive the figure of three million pieces of content a day from Zuckerberg's reference in this open letter to Facebook reviewing "over 100 million pieces of content every month."

[8] Mark Zuckerberg, "A Blueprint for Content Governance Enforcement," Facebook, November 15, 2018, https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/.

[9] Benjamin Mullin, "In a Policy Change, Facebook Will Allow More Newsworthy Graphic Content," *Poynter*, October 21, 2016, https://www.theguardian.com/technology/2016/sep/09/facebook-reinstates-napalm-girl-photo.

[10] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech," *Harvard Law Review*, April 10, 2018, https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/.

[11] Meredith Bennett-Smith, "Facebook Vows to Crack Down on Rape Joke Pages After Successful Protest, Boycott," *Huffington Post*, May 29, 2013, https://www.huffpost.com/entry/facebook-rape-jokes-protest_n_3349319.

[12] Charlie Warzel, "'A Honeypot for Assholes': Inside Twitter's 10-Year Failure to Stop Harassment," *BuzzFeed News*, August 11, 2016, https://www.buzzfeednews.com/article/charliewarzel/a-honeypot-for-assholes-inside-twitters-10-year-failure-to-s.

[13] Nitasha Tiku and Casey Newton, "Twitter CEO: 'We Suck at Dealing With Abuse,'" *The Verge*, April 4, 2015, https://www.theverge.com/2015/2/4/7982099/twitter-ceo-sent-memo-taking-personal-responsibility-for-the.

[14] Mark Zuckerberg, "Building Global Community," supra note 7.

[15] Deepa Seetharaman and Joshua Jamerson, "After Posting Violent Videos, Facebook Will Add 3,000 Content Moderators," *The Wall Street Journal*, May 3, 2017, https://www.wsj.com/articles/zuckerberg-says-facebook-will-add-3-000-people-to-review-content-after-violent-posts-1493822842.

[16] Ariana Tobin, Madeleine Varner, and Julia Angwin, "Facebook's Uneven Enforcement of Hate Speech Rules Allows Vile Posts to Stay Up," *ProPublica*, December 28, 2017, https://www.propublica.org/article/facebook-enforcement-hate-speech-rules-mistakes.

[17] David Kirkpatrick, The Facebook Effect, Simon & Schuster, 2010, https://www.simonandschuster.com/books/The-Facebook-Effect/David-Kirkpatrick/9781439102121.

[18] "Content Moderators Reach Class Settlement with Facebook for $52 Million and Workplace Improvements," *PR Newswire*, May 12, 2020, https://www.prnewswire.com/news-releases/content-moderators-reach-class-settlement-with-facebook-for-52-million-and-workplace-improvements-301058100.html.

[19] Madhumita Murgia, "Facebook Content Moderators Required to Sign PTSD Form," *Financial Times*, January 26, 2020, https://www.ft.com/content/98aad2f0-3ec9-11ea-a01a-bae547046735; Casey Newton, "YouTube Moderators Are Being Forced to Sign a Statement Acknowledging the Job Can Give Them PTSD," *The Verge*, January 24, 2020, https://www.theverge.com/2020/1/24/21075830/youtube-moderators-ptsd-accenture-statement-lawsuits-mental-health.

[20] Sowmya Karunakaran and Rashmi Ramakrishnan, "Testing Stylistic Interventions to Reduce Emotional Impact of Content Moderation Workers," Conference on Human Computation and Crowdsourcing, 2019, https://www.aaai.org/ojs/index.php/HCOMP/article/view/5270/5122.

[21] Munsif Vengattil and Paresh Dave, "Facebook Contractor Hikes Pay for Indian Contract Reviewers," August 19, 2019, https://www.reuters.com/article/us-facebook-reviewers-wages/facebook-contractor-hikes-pay-for-indian-content-reviewers-idUSKCN1V91FK.

[22] Joshua Brustein, "Facebook Grappling with Employee Anger Over Moderator Conditions," supra note 4.

[23] Casey Newton, "The Terror Queue," *The Verge*, December 16, 2019, https://www.theverge.com/2019/12/16/21021005/google-youtube-moderators-ptsd-accenture-violent-disturbing-content-interviews-video. Links to earlier articles in the series are contained within this one.

[24] Casey Newton, "A Facebook Content Moderation Vendor Is Quitting the Business After Two Verge Investigations," *The Verge*, October 30, 2019, https://www.theverge.com/2019/10/30/20940956/cognizant-facebook-content-moderation-exit-business-conditions-investigation.

[25] Adam Taylor, "The Big Questions for Mark Zuckerberg on Facebook's Role in Burma," *The Washington Post*, April 10, 2018, https://www.washingtonpost.com/news/worldviews/wp/2018/04/10/the-big-questions-for-mark-zuckerberg-on-facebooks-role-in-burma/; Vindu Goel and Shaikh Azizur Rahman, "When Rohingya Refugees Fled to India, Hate on Facebook Followed," *The New York Times*, June 14, 2019, https://www.nytimes.com/2019/06/14/technology/facebook-hate-speech-rohingya-india.html.

[26] Vindu Goel and Shaikh Azizur Rahman "When Rohingya Refugees Fled to India, Hate on Facebook Followed," supra note 25.

Endnotes (continued)

27 "Megaphone for Hate," Avaaz, October 2019, https://avaazpress.s3.amazonaws.com/FINAL-Facebook%20in%20Assam_Megaphone%20for%20hate%20-%20Compressed%20(1).pdf.

28 "Facebook India—Towards a Tipping Point of Violence," Equality Labs, June 2019, https://www.equalitylabs.org/facebookindiareport.

29 Miranda Sissons and Alex Warofka, "An Update on Facebook's Human Rights Work in Asia and Around the World," Facebook, May 12, 2020, https://about.fb.com/news/2020/05/human-rights-work-in-asia/.

30 Amanda Taub and Max Fisher, "Where Countries Are Tinderboxes and Facebook Is a Match," *The New York Times*, April 21, 2018, https://www.nytimes.com/2018/04/21/world/asia/facebook-sri-lanka-riots.html; Michael Safi, "Sri Lanka Accuses Facebook Over Hate Speech After Deadly Riots," *The Guardian*, March 14, 2018, https://www.theguardian.com/world/2018/mar/14/facebook-accused-by-sri-lanka-of-failing-to-control-hate-speech.

31 Miranda Sissons and Alex Warofka, "An Update on Facebook's Human Rights Work in Asia and Around the World," supra note 29.

32 Newley Purnell, "Sri Lankan Islamist Called for Violence on Facebook Before Easter Attacks," *The Wall Street Journal*, April 30, 2019, https://www.wsj.com/articles/sri-lankan-islamist-called-for-violence-on-facebook-before-easter-attacks-11556650954.

33 Miranda Sissons and Alex Warofka, "An Update on Facebook's Human Rights Work in Asia and Around the World," supra note 29.

34 Id.

35 Id.

36 Id.

37 Fanny Potkin, "Indonesia Curbs Social Media, Blaming Hoaxes for Inflaming Unrest," *Reuters*, May 22, 2019, https://www.reuters.com/article/indonesia-election-socialmedia/indonesia-curbs-social-media-blaming-hoaxes-for-inflaming-unrest-idUSL4N22Y2UP.

38 Abdi Latif Dahir, "Nobel Peace Laureate Says Social Media Sows Hate in Ethiopia," *The New York Times*, December 10, 2019, https://www.nytimes.com/2019/12/10/world/africa/nobel-peace-abiy-ahmed.html.

39 Simon Marks, "67 Killed in Ethiopia Unrest, but Nobel-Winning Prime Minister Is Quiet," *The New York Times*, October 25, 2019, https://www.nytimes.com/2019/10/25/world/africa/ethiopia-protests-prime-minister.html. The updated fatality figure of 86 is cited in Ethiopian news publications—e.g., "Bodyguards Recalled by Federal Police Are Staying With Me: Jawar Mohammed," *Ezega News*, February 8, 2020, https://www.ezega.com/News/NewsDetails/7731/Bodyguards-Recalled-by-Federal-Police-are-Staying-with-Me-Jawar-Mohammed.

40 Dawit Endeshaw, "Ethiopia's Army Chief, Three Others Killed in Failed Regional Coup," *Reuters*, June 23, 2019, https://www.reuters.com/article/us-ethiopia-security/ethiopias-army-chief-three-others-killed-in-failed-regional-coup-idUSKCN1TO0CR.

41 Dawit Endeshaw, "Ethiopia Passes Law Imposing Jail Terms for Internet Posts that Stir Unrest," *Reuters*, February, 13, 2020, https://www.reuters.com/article/us-ethiopia-politics/ethiopia-passes-law-imposing-jail-terms-for-internet-posts-that-stir-unrest-idUSKBN2071PA.

42 Paul M. Barrett, "Disinformation and the 2020 Election: How the Social Media Industry Should Prepare," New York University Stern Center for Business and Human Rights, September 2019, https://bhr.stern.nyu.edu/tech-disinfo-and-2020-election.

43 Lauren Weber and Deepa Seetharaman, "The Worst Job in Technology: Staring at Human Depravity to Keep It Off Facebook," *The Wall Street Journal*, December 27, 2017, https://www.wsj.com/articles/the-worst-job-in-technology-staring-at-human-depravity-to-keep-it-off-facebook-1514398398.

44 Casey Newton, "What Tech Companies Should Do About Their Content Moderators' PTSD," *The Verge*, January 28, 2020, https://www.theverge.com/interface/2020/1/28/21082642/content-moderator-ptsd-facebook-youtube-accenture-solutions.

45 Mark Zuckerberg, "The Internet Needs New Rules. Let's Start in These Four Areas," *The Washington Post*, March 30, 2019, https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html; Monika Bickert, "Charting a Way Forward: Online Content Regulation," Facebook, February 2020, https://about.fb.com/wp-content/uploads/2020/02/Charting-A-Way-Forward_Online-Content-Regulation-White-Paper-1.pdf.

46 "The International Fact-Checking Network," Pointer, undated, https://www.poynter.org/ifcn/.

NYU Stern Center for Business and Human Rights
Leonard N. Stern School of Business
44 West 4th Street, Suite 800
New York, NY 10012
+1 212-998-0261
bhr@stern.nyu.edu
bhr.stern.nyu.edu

NYU | STERN

Center for Business
and Human Rights