Helen He
FRSEMR 70Z Regulating Online Conduct
Final Paper
December 19, 2019

The People Behind the Screen: A Look Into the Online Content Moderation Industry

I. Introduction

Let's start with a hypothetical. Imagine that you return home from a long day at work—back to back meetings, tense lunch conversations, and a traffic-heavy evening commute. You grab a snack from your fridge, sit down at your kitchen counter, open up your laptop, and type in "www.facebook.com," almost as if out of muscle memory. What do you see once the page loads? Instead of the usual pictures of your friends, notifications of nearby events, and memes from the community group you are a member of, your feed is ridden with blatant political propaganda, terrorist videos, and child pornography. As you scroll along, you see videos of beheadings, pictures of extreme nudity, and phrases with extremely offensive and defamatory language.

It might be hard to imagine your reaction to such a sight on your Facebook page. That's because, for the most part, none of us have experienced anything like this. Our social media feeds are pruned to only exhibit the posts we want to see, or that society generally deems as "okay" to share (sometimes to an extreme extent and with far-reaching negative consequences, but that is a topic for another paper). The way this happens may seem natural—accidental, even—but that could not be further from the truth. This paper will shed light on the online content moderators behind every post we see—describing how the process of online content

moderation works today, examining the labor issues associated with the industry, and proposing potential solutions to these issues.

II. Background: How the Process Works Today

While the statistics of the industry are not fully known, some experts estimate there to be tens of thousands of people working as content moderators around the world.[1] Each moderator can go through up to 8,000 posts a day, and spend as little as eight seconds on each post on average.[2]

Most U.S. tech companies outsource this work to external firms—with common partners being TaskUs and Open Access BPO—who provide content moderation services.[3] These firms largely contract workers abroad, with many hiring laborers from the Philippines—partly because of the country's cultural ties with the U.S. following their period of American colonization in the 1900s. These cultural overlaps can be helpful in determining what users in America will and will not find to be offensive.[4] At the same time, though, given the increasing demand of tech companies for these content moderators, a fraction of the work is also contracted domestically as well—with young college graduates making up most of the population of content moderators in the U.S. today.[5]

---

[1] Weber, Lauren. "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?ns=prod/accounts-wsj.
[2] Ibid; Trenholm, Richard. "The Cleaners documentary crawls into the scary side of Facebook." *CNet.* November 12, 2018. https://www.cnet.com/news/the-cleaners-sundance-documentary-review-dirt-on-social-media-fake-news/.
[3] Chen, Adrian. "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed." *Wired*. October 23, 2014. https://www.wired.com/2014/10/content-moderation/.
[4] Ibid.
[5] Ibid.

There are two main types of moderation in the industry today: active moderation and reactive moderation—with the former involving the moderation of every single post in real-time, and the latter the examination of only the posts flagged by users.[6] Particularly with reactive moderation, posts marked for review are often assigned to specific teams of workers, who are grouped by language and cultural proficiency, in order to ensure that the moderator working on a certain post has a sufficient understanding of the post at hand.[7] Additionally, many companies operate with two tiers of moderation, where basic and simpler decisions are outsourced abroad while more complex decisions, requiring more cultural familiarity, are assigned to workers within the U.S.[8]

The general process for a content moderation team may, then, look like the following: a post is reported by a user, then it is "automatically routed to a content review team based on language or the type of violation." Each moderator is then assigned a list of reported posts to evaluate—where each is also accompanied by the comments associated with it in order to provide reviewers the context needed in order to make the most accurate decision.[9] The reviewers then spend their workdays going through each post one by one, referencing the guidelines given to them by their firm or the tech company itself, and then deciding whether to

---

[6] Ibid.
[7] Silver, Ellen. "Hard Questions: Who Reviews Objectionable Content on Facebook — And Is the Company Doing Enough to Support Them?" *Facebook*. July 26, 2018. https://about.fb.com/news/2018/07/hard-questions-content-reviewers/.
[8] Chen, Adrian. "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed." *Wired*. October 23, 2014. https://www.wired.com/2014/10/content-moderation/.
[9] Silver, Ellen. "Hard Questions: Who Reviews Objectionable Content on Facebook — And Is the Company Doing Enough to Support Them?" *Facebook*. July 26, 2018. https://about.fb.com/news/2018/07/hard-questions-content-reviewers/.

delete or ignore it before moving on to the next one.[10] And, to ensure accuracy, these firms will often randomly choose a number of posts to check, or "audit."[11]

III. Mental Health Issues

The world of content moderation raises a number of concerns regarding employees' wellbeing—including concerns about mental health, as well as labor issues relating to the strict, inhumane working conditions common to the industry.

Given the inherently harsh and gruesome nature of the work, it's no surprise that content moderation takes a heavy emotional toll on reviewers. These laborers are exposed to violent and disturbing content at an extremely high volume, often to the point where they begin feeling numbed.[12] One moderator who was interviewed in the 2018 documentary *The Cleaners* described, "I've seen hundreds of beheadings."[13] Another reviewer explains, "If someone was uploading animal abuse, a lot of the time it was the person who did it. He was proud of that… And seeing it from the eyes of someone who was proud to do the fucked-up thing, rather than news reporting on the fucked-up thing—it just hurts you so much harder, for some reason. It just gives you a much darker view of humanity."[14]

---

[10] Chen, Adrian. "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed." *Wired*. October 23, 2014. https://www.wired.com/2014/10/content-moderation/.
[11] Silver, Ellen. "Hard Questions: Who Reviews Objectionable Content on Facebook — And Is the Company Doing Enough to Support Them?" *Facebook*. July 26, 2018. https://about.fb.com/news/2018/07/hard-questions-content-reviewers/.
[12] Chen, Adrian. "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed." *Wired*. October 23, 2014. https://www.wired.com/2014/10/content-moderation/.
[13] Trenholm, Richard. "The Cleaners documentary crawls into the scary side of Facebook." *CNet.* November 12, 2018. https://www.cnet.com/news/the-cleaners-sundance-documentary-review-dirt-on-social-media-fake-news/.
[14] Chen, Adrian. "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed." *Wired*. October 23, 2014. https://www.wired.com/2014/10/content-moderation/.

These emotional and mental health issues extend far beyond the workplace itself. Seeing this type of content every day can lead employees to develop issues like insomnia, anxiety, PTSD, and paranoia. In a lawsuit filed against Microsoft, one reviewer describes how he "found it difficult at times to be near computers or his own son."[15] Others moderators report becoming intensely paranoid, and beginning "to suspect the worst of people they meet in real life, wondering what secrets their hard drives might hold." One victim of this became "so suspicious that they no longer [left] their children with babysitters" and, as a result, would "sometimes miss work because they [couldn't] find someone they trust[ed] to take care of their kids."[16] For some moderators, the emotional toll becomes so unbearable that they choose to end their own lives. These issues from work can also influence employees' ideological beliefs, with some reviewers beginning to develop fringe views as a result of the conspiracy videos they frequently watch. As one journalist visiting a content moderation company describes, "One auditor walks the floor promoting the idea that the Earth is flat. A former employee told me he has begun to question certain aspects of the Holocaust. Another former employee, who told me he has mapped every escape route out of his house and sleeps with a gun at his side, said: 'I no longer believe 9/11 was a terrorist attack.'"[17] Finally, exposure to this disturbing content can even influence moderators' personal lives—where, for instance, watching extensive amounts of pornography

---

[15] Weber, Lauren. "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?ns=prod/accounts-wsj.

[16] Chen, Adrian. "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed." *Wired*. October 23, 2014. https://www.wired.com/2014/10/content-moderation/.

[17] Newton, Casey. "The Trauma Floor." *The Verge*. February 25, 2019. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

has left some avoiding sexual intimacy with their partners and others experiencing supercharged sex drives.[18]

In response to these heavy, emotional disturbances, many content moderators develop unhealthy coping mechanisms. One employee describes the industry as "an environment where workers cope by telling dark jokes about committing suicide, then smoke weed during breaks to numb their emotions."[19] While some workers turn to drugs and alcohol to help them handle the emotions of their job, others resort to telling offensive jokes. One article describes, "the workplace was rife with pitch-black humor. Employees would compete to send each other the most racist or offensive memes… in an effort to lighten the mood."[20] One employee who was interviewed for the article said, "I had to watch myself when I was joking around in public. I would accidentally say [offensive] things all the time — and then be like, *Oh shit, I'm at the grocery store. I cannot be talking like this*."[21] Lastly, others resort to sexual activity in desperation "for a dopamine rush amid the misery" of their job, with many reviewers found "having sex insides stairwells and a room reserved for lactating mothers." Employees in the industry have started calling these incidences "trauma bonding," reflective of the deeply disturbing experiences these moderators undergo while on the job.[22]

Further compounding the mental harm of this profession is isolation. Many content reviewers are required to sign non-disclosure agreements (NDAs) which prohibit them from

---

[18] Chen, Adrian. "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed." *Wired*. October 23, 2014. https://www.wired.com/2014/10/content-moderation/.
[19] Newton, Casey. "The Trauma Floor." *The Verge*. February 25, 2019. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.
[20] Ibid.
[21] Ibid.
[22] Ibid.

speaking about their work with loved ones, and sometimes even with other employees.[23] These NDAs are supposedly meant to "protect employees from users who may be angry about a content moderation decision and seek to resolve it with" the reviewers directly. Instead, however, these non-disclosure restrictions further isolate employees in their emotional distress and allow their employers to get away with harsh and unfair working conditions.[24]

While mental health services for moderators do exist, they are nowhere near the level of effectiveness and availability needed to fully support the mental challenges of this job. For example, while Cognizant—a content moderation firm—has a hired counselor, they are only available to employees for part of the day.[25] And, even though YouTube has also designated psychologists to support moderators, employees report that they often don't know how to access these resources.[26] Even if workers are able to meet with counselors, many describe them as "largely passive" and reliant "on workers to recognize the signs of anxiety and depression."[27] Worse yet, a number of psychologists are not fully and genuinely invested in supporting struggling employees, but rather seem more concerned with business goals of appeasing workers and convincing them to not quit the job.[28] One such tactic involves diagnosing moderators with "post-traumatic growth" instead of PTSD, in order to claim that they are trauma victims who

---

[23] Ibid; Chen, Adrian. "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed." *Wired*. October 23, 2014. https://www.wired.com/2014/10/content-moderation/.
[24] Newton, Casey. "The Trauma Floor." *The Verge*. February 25, 2019. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.
[25] Ibid.
[26] Chen, Adrian. "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed." *Wired*. October 23, 2014. https://www.wired.com/2014/10/content-moderation/.
[27] Newton, Casey. "The Trauma Floor." *The Verge*. February 25, 2019. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.
[28] Ibid.

"emerge from the experience feeling stronger than before."[29] Finally, these services—if any at all—end immediately after a moderator quits or is fired, meaning that these workers receive no support following their employment. This is especially concerning given the common lasting negative effects of these mental health issues.[30] Ultimately, though, even if these mental health services were effective and available, counseling itself cannot eliminate the disturbances and trauma experienced by these moderators. As one psychologist explains, "It's like PTSD… There is a memory trace in their mind." [31]

IV. Other Labor Issues

In addition to the mental health challenges ingrained in their work, these content moderators also face extremely harsh working conditions imposed by their employers. Some have described their workplace as a "digital sweatshop," while others characterize it as a "Big Brother environment."[32] Because these employees are contracted gig workers as opposed to full time employees, there are few labor laws that protect them or govern their work. This creates a class of "invisible workers" who often earn less than the legal minimum wage, receive very few benefits, and have almost no job security.[33]

Content moderators are paid extremely poorly compared to their full-time counterparts within the same company. For example, within Facebook, moderators in Arizona make only

---

[29] Ibid.
[30] Ibid.
[31] Chen, Adrian. "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed." *Wired*. October 23, 2014. https://www.wired.com/2014/10/content-moderation/.
[32] Trenholm, Richard. "The Cleaners documentary crawls into the scary side of Facebook." *CNet.* November 12, 2018. https://www.cnet.com/news/the-cleaners-sundance-documentary-review-dirt-on-social-media-fake-news/; Price, Rob. "Facebook moderators are in revolt over 'inhumane' working conditions that they say erodes their 'sense of humanity.'" *Business Insider*. February 15, 2019. https://www.businessinsider.com/facebook-moderators-complain-big-brother-rules-accenture-austin-2019-2.
[33] "Ghost Work." *Ghost Work*. Accessed December 19, 2019. https://ghostwork.info/ghost-work/.

$28,800 a year, while the average Facebook employee is paid $240,000.[34] This disparity is only exacerbated when we compare salaries between countries, with even a veteran Filipino moderator making less in a day than a brand-new American moderator in an hour.[35] Worse yet, these outsourced workers are often very poor and thus feel pressure to remain in their jobs in order to support their family, even if the work damages their own wellbeing.[36]

These moderators also face strict expectations for the accuracy of the work they do—despite the inherently ambiguous nature of determining whether a post is appropriate or not. Workers in this industry are evaluated on accuracy, where this accuracy is judged only on agreement with a supervisor.[37] Employees are expected to maintain a 98% accuracy rate and, if they fall shy of this threshold, they often risk losing their job.[38] This high-pressure environment also generates hostility between workers, where the quality assurance workers who are responsible for confirming the accuracy of certain decisions often feel threatened by previously fired employees who return seeking vengeance. For example, one article describes the experience of a quality assurance worker who "would sometimes return to his car at the end of a work day to find moderators waiting for him. Five or six times over the course of a year, someone would attempt to intimidate him into changing his ruling." In the worker's own words,

---

[34] Newton, Casey. "The Trauma Floor." *The Verge*. February 25, 2019. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

[35] Chen, Adrian. "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed." *Wired*. October 23, 2014. https://www.wired.com/2014/10/content-moderation/.

[36] Trenholm, Richard. "The Cleaners documentary crawls into the scary side of Facebook." *CNet.* November 12, 2018. https://www.cnet.com/news/the-cleaners-sundance-documentary-review-dirt-on-social-media-fake-news/.

[37] Newton, Casey. "The Trauma Floor." *The Verge*. February 25, 2019. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

[38] Price, Rob. "Facebook moderators are in revolt over 'inhumane' working conditions that they say erodes their 'sense of humanity.'" *Business Insider*. February 15, 2019. https://www.businessinsider.com/facebook-moderators-complain-big-brother-rules-accenture-austin-2019-2.

"They would confront me in the parking lot and tell me they were going to beat the shit out of me… There wasn't even a single instance where it was respectful or nice. It was just, *You audited me wrong! That was a boob! That was full areola, come on man!*" As a result, this worker now brings a gun to work each day because he feels threatened and unsafe.[39]

Finally, employers are also incredibly strict on content moderators' efficiency at work. Managers monitor these employees remotely, and if they notice even just a few minutes of inactivity, they message the worker to ask why they aren't working.[40] Lastly, breaks and "wellness time" are both limited and managed closely—with employers dictating specifically what employees can and can't do during these short periods of time.[41]

V. Potential Solutions

Given the double bind of the inherently disturbing nature of content moderation and the necessity of this job in ensuring the well-being of the online society as a whole, addressing the issues of the content moderation industry is admittedly a challenge. There are, however, a number of ways the damages of this work can be minimized and the employees in the industry can be better supported.

First, more money should be invested in improving the working conditions of the content moderation industry. Whether this means improving the efficacy and accessibility of mental

---

[39] Newton, Casey. "The Trauma Floor." *The Verge*. February 25, 2019. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

[40] Weber, Lauren. "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?ns=prod/accounts-wsj.

[41] Newton, Casey. "The Trauma Floor." *The Verge*. February 25, 2019. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

health services, hiring higher numbers of moderators in order to reduce the workload on any individual worker, or providing higher compensatory wages to these moderators to support them for their sacrifices, these monetary investments seem necessary for the sake of promoting fair labor practices and fundamental humane integry.

Secondly, a shift to relying on technology and artificial intelligence (AI) also seems to be an appealing solution. This would allow all humans to step (further) away from the dark side of the internet, and instead have machines take care of the filtering. And, the good news is that companies already recognize this as potential solution and are currently investing in developing this technology. Nathaniel Gleicher, Head of Cybersecurity Policy at Facebook, wrote in a blog post, "By using technology like machine learning, artificial intelligence and computer vision, we can proactively detect more bad actors and take action more quickly." He writes of the current state of the technology, "We took down 837 million pieces of spam and 2.5 million pieces of hate speech and disabled 583 million fake accounts globally in the first quarter of 2018 — much of it before anyone reported the issue to Facebook."[42] More specifically, they have developed features that identify posts, pictures, comments, or other content that is likely to violate their community standards—such as through Instagram's online-bullying detecting feature which "notifies people when their captions on a photo or video may be considered offensive, and gives them a chance to pause and reconsider their words before posting."[43]

Ultimately, though it's still difficult to imagine a time where AI will be as good as humans at detecting the nuances of these decisions—especially when it comes to considering the broader contexts of reported posts. Regardless, the negative impacts of the current system of

---

[42] Gleicher, Nathaniel. "Removing Bad Actors From Facebook." *Facebook*. June 26, 2018. https://about.fb.com/news/2018/06/removing-bad-actors-from-facebook/.
[43] Ibid.

content moderation are significant enough to warrant serious consideration of sacrificing some free speech in order to make the process more automated. This system has already begun to be adopted in China, where "many Chinese social media platforms preemptively prevent people from posting content that contains certain words. Others automatically delete posts with those words."[44] It may be worth it to turn much of this work over to machines—even if it means accepting more frequent mistakes or removing content that could otherwise be left up—so that humans can be spared from the emotional (and physical) abuse inherent to challenging work.

[44] Myers, Sarah. "Mistreated moderators and the pervasive violence of the internet." *The Stanford Daily*. March 6, 2019. https://www.stanforddaily.com/2019/03/06/me-ll-mistreated-moderators-and-the-pervasive-violence-of-the-internet/ .

VI. Works Cited

Bickert, Monika. "Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process." *Facebook*. April 24, 2018. https://about.fb.com/news/2018/04/comprehensive-community-standards/.

Bishop, Bryan. "The Cleaners is a riveting documentary about how social media might be ruining the world." *The Verge.* January 21, 2018. https://www.theverge.com/2018/1/21/16916380/sundance-2018-the-cleaners-movie-review-facebook-google-twitter .

Chen, Adrian. "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed." *Wired*. October 23, 2014. https://www.wired.com/2014/10/content-moderation/.

"Ghost Work." *Ghost Work*. Accessed December 19, 2019. https://ghostwork.info/ghost-work/.

Gleicher, Nathaniel. "Removing Bad Actors From Facebook." *Facebook*. June 26, 2018. https://about.fb.com/news/2018/06/removing-bad-actors-from-facebook/.

Myers, Sarah. "Mistreated moderators and the pervasive violence of the internet." *The Stanford Daily*. March 6, 2019. https://www.stanforddaily.com/2019/03/06/me-ll-mistreated-moderators-and-the-pervasive-violence-of-the-internet/ .

Newton, Casey. "The Trauma Floor." *The Verge*. February 25, 2019. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

"Our Progress on Leading the Fight Against Online Bullying." *Facebook*. December 16. 2019. https://about.fb.com/news/2019/12/our-progress-on-leading-the-fight-against-online-bullying/.

Price, Rob. "Facebook moderators are in revolt over 'inhumane' working conditions that they say erodes their 'sense of humanity.'" *Business Insider*. February 15, 2019. https://www.businessinsider.com/facebook-moderators-complain-big-brother-rules-accenture-austin-2019-2.

Silver, Ellen. "Hard Questions: Who Reviews Objectionable Content on Facebook — And Is the Company Doing Enough to Support Them?" *Facebook*. July 26, 2018. https://about.fb.com/news/2018/07/hard-questions-content-reviewers/.

Trenholm, Richard. "The Cleaners documentary crawls into the scary side of Facebook." *CNet.* November 12, 2018. https://www.cnet.com/news/the-cleaners-sundance-documentary-review-dirt-on-social-media-fake-news/.

Weber, Lauren. "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?ns=prod/accounts-wsj.

Wong, Julia Carrie. "Facebook contractors faced Christmas ultimatum: accept wage offer or lose jobs." The Guardian. December 20, 2018. https://www.theguardian.com/technology/2018/dec/20/facebook-contractors-filter-digital-labor-dispute-christmas.