The Effects of Facebook's Response to Content Moderators

A Research Paper
in STS 4600
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By

Cory Ayers

April 10th, 2020

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signed: Cory Ayers Date: 4/10/20

Introduction

Everyone that has surfed the internet has seen something, whether it be on Facebook, Google, or another site, that disturbed them or made them uncomfortable. This content can come in many forms, from graphic images to videos to online harassment. Now, imagine seeing these things every second for eight hours each workday. The content moderators at Facebook are tasked with this problem, protecting one of the world's most used social media sites from the most disturbing content imaginable. The reason these sites seem relatively safe is due to the moderators who work tirelessly to improve the lives of others. Unfortunately, extended exposure to this form of media, especially by moderators of these sites, can have long-term negative effects on their mental health and everyday lives.

Recently, information has come out about Facebook's role in creating these hostile work environments and how they have done little to change issues they knew were present (Newton, 2019, p. 5). Content moderators complain they do not receive enough psychological services, and that they are not allowed to talk to friends due to strict confidentiality agreements (Cabato et al., 2019, p. 2). The public has gained information through former employees breaking non-disclosure agreements (NDAs), but the secrecy surrounding the subject at different offices around the world means there may be much more going on (Newton, 2019, p. 4). It can be so extreme, some workers now claim to sleep every night with a gun in their hands, one even sweeps his house every morning in fear that someone may be inside (Newton, 2019, p. 33). The health of each of these individuals is vitally important to all that use social media, because these are the people that keep sites much more user friendly and appropriate when we use them every day. They are the reason terrorist organizations are not something you see pop up on your news feed and videos of people dying in them are a rarity for most users. These moderators are not

displayed as super users, with special powers over the rest of the site, but instead are hidden all around the world, with little ability to tell others about what they are actually going through (Roberts, 2019, p. 2). Without these employees, parents would not let their children go online, or let high schoolers download social media, for the fear of violent, traumatizing, hateful content likely to surface on their feeds.

Through my research, I plan on narrowing the knowledge gap between what moderators are experiencing and the general perception by the public. I have looked deeper into the conditions they work under, the stress their managers cause, and the horrible damage this job has had on thousands of former and current employees. This came from a variety of sources, including books that have studied the lives of moderators for close to a decade, interviews by real people with real moderators, and testimonies from companies and court filings. Finally, I examined Facebook's response to the recent public backlash over these extreme accusations, and am looking to determine if the conditions are finally improving or not. This will be important for the future of thousands of current contractors, as well as the future of social media as we know it. If Facebook does not improve these situations, contracting companies will be forced to either quit working with Facebook, or continue to attempt hiding the hostile working conditions these employees are living under every single day.

Background

The idea of moderation is as old as the internet is, and has been going on for nearly four decades, whether you have noticed it or not (Roberts, 2019, p. 1). Every social media platform must moderate, it is an inherently vital part of its survival. The goal of the majority of these sites is to facilitate conversation, host users, moderate content, and profit from said users (Gillespie,

2018, p. 15). This is where the problem has begun, many social media platforms have started caring about profits over the well being of their workers and users. This issue has only come to light in the past decade or so because these platforms have grown to a scale where moderation of individual posts is virtually impossible, where thousands of new posts show up each second worldwide (Gillespie, 2018, p. 7). When it comes to this, companies have to make very tough decisions, how much to moderate so people are not viewing inappropriate material, but not too much that they feel their free speech is being taken from them. This is what Facebook is currently going through, an attempt to keep their users happy and still using the site while also hiding the effects it is having on their thousands of moderators worldwide.

The atrocities that surround content moderation globally have been going on for decades, but are only recently surfacing in public light. In Asia, one employee discussed how skyscrapers give him constant flashbacks to all of the suicides he has reviewed (Dwoskin et al., 2019, p. 18). Even though he has quit his job, constant therapy and special assistance is still needed to help him get through everyday life (Cabato et al., 2019, p. 18). These results are clearly serious and significant, as seen through studies done by UCF and Penn State professors. They have examined how negative online experiences can lead to symptoms of post-traumatic stress disorder (PTSD) in victims, including looking at explicit/sexual content and cyberbullying (McHugh et al., 2018, p. 1). Having these interactions every day might take an even worse toll on workers, as shown in these examples.

Looking at Facebooks benefits and policies surrounding this situation, it appears they are not meant to help workers feel safe and cared for. Many former employees stated that they were staring at content non-stop and that they were strongly encouraged to examine a certain amount of content each day (Newton, 2019, p. 16). Other employees complain of a lack of counseling

available and that they are limited to very few breaks during the workweek (Newton, 2019, p. 23). Another noted that they had to log every hour of their days at work, including when they were in the bathroom (Gilbert, D., 2020, p. 2). Part of this is a response to thresholds set by Facebook, that ask for moderators to judge posts correctly a stunning 95 percent of the time (Newton, 2019, p. 14). Managers at these offices also examine the amount of content each employee goes through, and their jobs could be at stake if they do not examine content fast enough (Newton, 2019, p. 16). This adds more stress to these intense jobs.

While Facebook helps set these policies, the workers are not actually in Facebook offices. Facebook, as well as other tech giants, use contracting companies, such as Accenture and Cognizant, to hire these employees to further protect themselves from the liabilities of content moderator health (Cabato et al., 2019, p. 6). In the past few years, plenty of information has come out to the public regarding this issue, largely due to former employees breaking NDA agreements. These former employees still felt the long-term effects of this psychologically straining job and felt it was their duty to speak up to protect future employees.

Finally, it appears as if Facebook contractors are fighting back even harder than before. New law suits have been filed, claiming the work done at these offices directly causes trauma and post-traumatic stress disorder. One of the most prominent cases comes out of Facebook Ireland, where multiple workers filed grievances against Facebook and their direct contracting employer, CPL (Woods, 2019, p. 1). One worker in the case claimed he had to review over 1000 postings every single night, and many of these posts have stuck with him for years after quitting his job. He noted these terrifying posts included, “footage of a woman being stoned to death, people being tortured with molten metal, and dogs being cooked alive.” (Woods, 2019, p. 1).

Examining the different players in this large system can help get a better understanding of what is really going on. Facebook is a multibillion-dollar corporation, with extreme amounts of power over their workers. If they need to, they can break away from contracting firms who are being targeted and hire new ones. When it comes to the workers of these contracting companies, there is little they can do to prevent what is happening to them. The jobs are extremely low paying, normally less than $30,000 a year (Woods, 2019, p. 3). Most workers take this job because it is one of the only ones they can find, and once they realize the exact work the job entails, it can be difficult to leave making money in exchange for their mental health and well-being. With all of these employees working under NDA agreements, it can be difficult for them to speak out without being sued for breaking contracts, making the world they live in far more secretive. Employees also point to this as a problem with coping mechanisms, they are not able to discuss the problems of their jobs with anyone.

Other literature has shown the problems of specific cases, such as law suits or the work of one specific firm, but not the entire situation and how it has been developing. I was not able to find any in-depth studies or literature reviews of this developing situation, partially because it has been so recent. Every month, more news has been coming out regarding this situation and I believe it is important to get the full scope of everything involving Facebook and these contracted content moderation workers.

Methods

I examined the responses of Facebook, contracting companies (such as Cognizant and Accenture), and the actual workers when looking at this case, as well as literature on the origins of content moderation and the secretive world these workers live in. When examining the

sociotechnical system here, the overall goal was an equal emphasis on the technical aspect of Facebook as well as the human aspect that involves all of these workers and their everyday lives. In this network of people, there is clearly a direct imbalance from the needs of workers and what the higher-ups at Facebook and contracting groups expect. I looked at this balance more and to see how Facebook is responding to the criticisms of imbalance in this system and how technological advancements (such as artificial intelligence) might be hindering their hopes of striking this balance. Along with this, I further examined how this system was created in the first place that lead to these developments for these workers.

I did this by finding as many current accounts of the situation as possible through news articles and interviews with current and former workers who have spoken out about problems they have had with Facebook. Along with this, I reviewed two books on the "Custodians of the Internet" and what exactly is going on behind the screen where so many users carelessly browse social media. I also looked into how Facebook is responding, from legal cases that are ongoing to their changes in policies around privacy, content moderation and their CEO, Mark Zuckerberg, speaking out about content moderation. My work will hopefully provide an in-depth review of this complex system and how it has evolved into what it is today. The result of this process is extremely important and necessary, the lives of over 30,000 current moderation workers are at stake in the battle between fair pay and compensation as well as mental health resources and breaks for workers.

Findings

Custodians of the Internet

To start my collection, I began reading about the origins of moderation on the internet and examples of why it is such a controversial and difficult subject. This began with a famous photo from the Vietnam war, that was taken off Facebook many times before eventually being allowed to stay on (Gillespie, 2018, p. 1). This photo helped illustrate just how diverse the content is that moderators go through and being reprimanded if they are not judging content correctly. With changes to their rule books coming in every single day, it can be extremely hard to keep track of exactly what constitutes a picture being removed, especially since this can change for any given post from week to week. Pictures such as this one dealing with the Vietnam war show just how complex some posts can be. Thousands of posts deal with violence and nudity, but are also culturally significant in some places around the world (Gillespie, 2018, p. 1). If images are removed for the wrong reason, workers can be reprimanded or even fired for making this decision that might require years of cultural knowledge to make an accurate decision on.

This is the root of the problem and how it all started for Facebook. The book, *Custodians of the Internet*, talks about how Facebook is "the world's most powerful editor" (Gillespie, 2018, p. 3). They get to make the decisions of what stays on and what leaves the site, and this can cause plenty of issues for users. In an attempt to fix this, they hire moderators, but Facebook cannot afford to pay millions of moderators a fair price to cover that much ground, they barely pay the moderators they do have a livable wage. Due to this, Facebook has continued to underpay its workers for years, while forcing them to work at harder paces under more difficult conditions as the company has continued to grow. This has led us to the current situation of today, where moderators are pressured daily to perform at high levels and take on more and more content with little counseling services or compensation.

Current Situation

When examining this entire situation and all of the intricacies that surround it, I found plenty of extremely interesting information and accounts of the situation currently surrounding content moderators and moderation as a whole at Facebook. The start of this involves the non-stop amount of testimonies from workers, describing horrific content they have witnessed on a day in, day out basis while working for content moderation contracting firms. They speak of horrific posts an average person could not bear to stomach, such as one woman who noted on her first day working, she witnessed someone being, "beaten to death with a plank of wood with nails in it and repeatedly stabbed" (Gilbert. D, 2091, p. 3). The trauma these employees face is not one off, it seems to encompass every firm in different countries worldwide.

Looking into Facebook and how they have responded to the situation, it appears they have done plenty to attempt to highlight their accomplishments in content moderation, while leaving out as much as they can about moderators. They claim to stop 99% of terrorist related posts before users can flag it, but studies by other outlets seem to believe this is completely untrue (Silver et al, 2018, p. 2). While Facebook is extremely good at removing and moderating posts by the top terrorist groups such as ISIS and Al-Qaeda, smaller groups seem to have free reign over the social media platform, even having links to their Facebook pages attempting to recruit new members on their own websites (Silver et al, 2018, p. 2). This seems to be in inherently flawed part of Facebooks current attempt to solve moderation issues with AI. Artificial intelligence is reasonably good at sifting through databases of past information and removing/flagging content based off of things that have already happened (Gillespie, 2018, p. 98). However, AI is unable to predict future problems and flag them when they finally occur. This is why longstanding terrorist groups and types of pornography are easily flagged, while new

terrorist groups, types of pornography and offensive language slip right through the cracks onto the screens of everyday users (Gillespie, 2018, p. 98).

Content moderation employees from all over have begun to sue Facebook, citing claims of PTSD and hostile work environments. While most of the cases are ongoing, they involve serious claims. These include that Facebook is well aware of the situations and is allowing contracting companies to get away with treating employees poorly. One contracting company actually starting making new employees sign an agreement acknowledging that their job could give them PTSD, and that employees could not do anything about it (Cellan-Jones, 2020, p. 1). One of the firms that is looking to sue Facebook is CPL, and its employees are going all out to get back at Facebook. Multiple sources from CPL cited needing anti-depressant drugs, along with alcohol, to get over the strain of their job (Gilbert, D., 2019, p. 3). Along with this, one of Facebook's higher up employees is also suing the company, saying they got Type II diabetes from looking at content that moderators flagged (Gilbert, D., 2019, p. 3). This type of powerful figure turning against them could pose a much larger threat for the multi-billion-dollar company. Paired with Mark Zuckerberg declining a petition to moderate content himself for one hour a day and calling the situation "a little overdramatic," this does not look good for Facebook (Wood, 2019, p. 2).

One of the apparent forces that shields Facebook from even more criticism is the use of contracting companies to do their content moderation work. By not having them be Facebook employees directly, they are able to pay them less, and have less apparent responsibility for them. Recently, one of their largest contractors, a company called Cognizant that operates out of Florida, announced it was cutting ties with Facebook (Holmes, 2019, p. 1). This began as a result of investigations into the firm earlier last year after multiple employees broke NDA's

about the firm and Facebook, to stand up for the thousands of employees who have stayed silent (Holmes, 2019, p. 2). Unfortunately, it will be extremely easy for Facebook to replace them with another contracting firm, as there are hundreds out there that will gladly accept this offer from Facebook.

By far the most interesting news that has come to light recently is Facebook's announcement that they are helping fund an entirely independent oversight board, spending over $130 million over the next six years to create and fund its everyday practices (Gilbert, B., 2020, p. 1). This board will review content moderation decisions and have the final say on what can stay and go when it comes to postings on Facebook, even having power where CEO Mark Zuckerberg himself cannot overturn rulings (Gilbert, B., 2020, p. 2). This decision comes as a direct result of the fast number of increasing users on Facebook in the past decade. In 2009, Facebook had only 12 content moderators for 120 million of users. Now they have over 30,000 moderators reviewing the everyday work of 2.2 billion users (Richter, 2019, p. 1). As this scale grows, the implications of this decision by Facebook are going to have clear and extreme impacts on the situations surrounding content moderation and moderators in the upcoming decade.

Significance

The findings of my work show how deeply rooted this issue is in social media and the internet. It appears that content moderation and its moderators have been enduring this type of behavior for far longer than most people have noticed, since it has only come to light in the form of mass social media. As media has continued to grow, it thankfully has brought many issues to light and is already starting to see some positive change. Companies are finally having to acknowledge the problems they have created and hopefully this will lead to meaningful help for

these workers. Facebook is being sued for certain wrongdoings and there is hope this will also lead to more compensation and safety for these workers in the future. While the current court case with Facebook and Irish workers for CPL is still ongoing, wins by the workers will hopefully push Facebook to acknowledge these problems more head on.

It has been difficult coming to the realization at the end of this study that it appears Facebook is not currently doing everything they could to help fix the problem of content moderators. Their sly tactics of removing the blame to a neutral organization that they fund is very apparent and upsetting to see they are trying to get away from any blame. Mark Zuckerberg has been heard many times saying the situation is overblown and that the media is wrong, but there are so many witnesses that say the exact opposite. It seems Facebook has done little investigation into what is actually going on at these contracting sites, probably because they do not want to know the specifics of the work conditions there. Nonetheless, it is important to continue to speak up for those whose voices are silenced, in hopes that change will continue to come in greater forms.

The continued progression of AI in the future is another promising form of helping these moderators and keeping the most graphic forms of content out of their sight. While perfection in artificial intelligence and machine learning is clearly impossible, it is believed it will get significantly better at stopping the most graphics images and videos from being viewed by anyone. The book, *Custodians of the Internet*, provided a few more useful ways that companies such as Facebook should explore for improving their content moderator problem. These included, "more transparency, better tools to block bad actors, better detection software, and more empathetic engagement with victims" (Gillespie, 2018, p. 198). If Facebook continued to be more transparent about the situation as a whole, more empathetic towards those who

wrongfully have posts blocked, and worked towards better detection software, the lives of moderators would be vastly improved.

Conclusion

The importance of the lives of these workers cannot go unseen. Thousands are suffering every single day, attempting to cope in silence with fears of losing their only source of income. It is so important to get their story out to the public to continue to create positive change surrounding the culture of social media moderation. While Facebook was the subject of my examination, there are hundreds of other technology and social media companies, including Google, Twitter, YouTube, and Instagram that employ these same tactics in regards to providing cheap labor where they can. The best way we can help those who are hurting is by giving them a voice and speaking out, in hopes that enough of us can make an impact on the world of content moderation.

Bibliography

Cabato, R., Dwoskin, E., & Whalen, J. (2019, July 25). Content moderators at YouTube, Facebook and Twitter see the worst of the web - and suffer silently. Retrieved from https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price/.

Carbone, C. (2019, November 1). 6000 jobs go as Facebook loses filter. Retrieved from https://www.news.com.au/technology/online/social/company-that-moderates-facebook-exits-amid-concerns-over-mental-health-of-moderators/news-story/3eb3c862852d75ff0b6d94abd251783a.

Carroll, J. M., McHugh, B. C., Rosson, M. B. & Wisniewski, P. (2018, October 2). When social media traumatizes teens: The roles of online risk exposure, coping, and post-traumatic stress. *Internet Research*, 28(5), 1169–1188. https://doi.org/10.1108/IntR-02-2017-0077

Cellan-Jones, R. (2020, January 25). Facebook and YouTube moderators sign PTSD disclosure. Retrieved from https://www.bbc.com/news/technology-51245616

Gilbert, B. (2020, February 16). Facebook is spending $130 million to create a 'Supreme Court' that can overrule Mark Zuckerberg - here's everything we know about it. Retrieved from https://www.businessinsider.com/facebook-moderation-independent-oversight-board-supreme-court-mark-zuckerberg-explained-2020-2

Gilbert, D. (2019, December 3). Bestiality, Stabbings, and Child Porn: Why Facebook Moderators Are Suing the Company for Trauma. Retrieved from https://www.vice.com/en_us/article/a35xk5/facebook-moderators-are-suing-for-trauma-ptsd

Gilbert, D. (2020, January 9). Facebook Is Forcing Its Moderators to Log Every Second of Their Days - Even in the Bathroom. Retrieved from https://www.vice.com/en_us/article/z3beea/facebook-moderators-lawsuit-ptsd-trauma-tracking-bathroom-breaks

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. New Haven: Yale University Press.

Holmes, A. (2019, October 31). The company behind Facebook's nightmarish moderation center in Florida will end its content moderation services. Retrieved from https://www.businessinsider.com/facebook-content-moderator-cognizant-cancels-contract-2019-10

Lapowsky, I. (2019, May 23). More Data on Content Moderation Won't Silence Facebook's Critics. Retrieved from https://www.wired.com/story/facebook-community-standards-report/.

Newton, C. (2019, February 25). The secret lives of Facebook moderators in America. Retrieved from https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Newton, C. (2019, June 19). Three Facebook moderators break their NDAs to expose a company in crisis. Retrieved from https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa.

Richter, F. (2019, March 13). Infographic: How Does Facebook Moderate Content. Retrieved from https://www.statista.com/chart/17302/facebook-content-moderator/#:~:text=Content

moderators on Facebook are,than regular salaried Facebook employees.&text=Now with a worldwide reach,offensive content on the internet.

Roberts, S. T. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media.* New Haven: Yale University Press.

Rosen, G. (2018, May 15). Facebook Publishes Enforcement Numbers for the First Time. Retrieved from https://newsroom.fb.com/news/2018/05/enforcement-numbers/.

Silver, V., & Frier, S. (2018, May 10). Terrorists Are Still Recruiting on Facebook, Despite Zuckerberg's Reassurances. Retrieved from https://www.bloomberg.com/news/articles/2018-05-10/terrorists-creep-onto-facebook-as-fast-as-it-can-shut-them-down

Wood, C. (2019, December 5). A former Facebook moderator is suing the firm, claiming that reviewing horrific material left him with PTSD. Retrieved from https://www.businessinsider.com/ex-facebook-mod-suing-claiming-disturbing-content-caused-ptsd-2019-12