Social Media Content Moderation: An Ethical Consideration

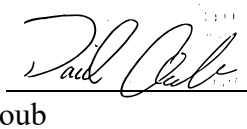
STS Research Paper
Presented to the Faculty of the
School of Engineering and Applied Sciences
University of Virginia

By

Daniel Ayoub

April 11, 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.

Signature:  Date 05/07/2020
Daniel Ayoub

Approved: _____ Date _____
Benjamin J. Laugelli, Assistant Professor, Department of Engineering and Society

Introduction

The Christchurch mosque shootings of March 2019 killed 51 people and injured 49 others. The attack was live-streamed on Facebook. “Including the views during the live broadcast, the video was viewed about 4,000 times in total before being removed from Facebook” (Rosen, 2019). For 12 minutes after the end of the live stream, the original video was available on the largest social media platform in the world for viewing or distribution. Surprisingly, Facebook’s artificial intelligence algorithms did not detect this video at upload, and thus it stayed on the site until it was later detected. Knowing that artificial intelligence is hardly sufficient alone, Facebook hires humans as commercial content moderators to filter out gruesome or offensive material uploaded to its site. Facebook’s use of humans as content moderators has typically been understood as a necessity towards the suppression of hate speech, global terrorism, and violence in order to protect humanity as a whole. However, this view fails to consider the moral dimensions inherent in asking a smaller subset of humans to view the same content being shielded from everyone else. To date, there is no prior literature discussing the ethics of the use of humans in these roles. If these moral dimensions continue to be neglected, we cease to consider the harm and possible injustices incurred on these employees, who do the work to keep our platforms safe behind the scenes. In this paper, I will argue that Facebook’s use of humans as content moderators is immoral, as to do so is to egregiously exploit humans as a means to a selfish end desired by Facebook. I will demonstrate how the use of humans in this way is disrespectful to their autonomies, damaging to their psychological well-beings, and negligent of their financial concerns. I will use Kantian ethics, which is a subset of duty ethics to evaluate the morality of Facebook’s decisions in this context.

Background

Social Media platforms have given individuals a voice of free speech, a voice that can be potentially heard around the world in a matter of seconds. With this voice, many people have brought to light vital initiatives to further the greater well-being of our society. People have used social media to advocate for a particular charity or cause, to enable others access to education, to support positive political movements, and to communicate with loved ones who may be far. However, despite all these positive uses, some have used social media platforms as a forum to spread hate speech, encourage violence, and popularize terrorism. In response, social media platforms have rushed to create complex algorithms that detect and block offensive and harmful content. In addition, they have hired thousands of human moderators who are responsible for the manual oversight and training of the algorithms. Chotiner of *The New Yorker* reports that "[m]ore than a hundred thousand people work as online content moderators, viewing and evaluating the most violent, disturbing, and exploitative content on social media" (Chotiner, 2019). Despite the great tasks before them, human moderators "are frequently relatively low-status and low-wage in relation to others in the tech industry and sometimes even in the same building" (S. Roberts, 2019). The decisions human moderators make in conjunction with the algorithms have massive implications, ones that can completely change a society for the better or worse. Alongside the social impact of content moderation, an ethical consideration arises from asking human moderators to watch and flag gruesome and potentially psychologically damaging material.

Literature Review

Several scholars have examined or considered the role of humans in social media content moderation. Some scholars bring to light the hidden logistics behind the use of humans in these roles: the working conditions, the real labor involved, and the wages received. Other works stress the emotional implications of the work on the employees. Lastly, some scholars analyze the sociopolitical, legal, and economical consequences of social media content moderation at large. Despite this wealth of information, scholars have yet to adequately determine if the use of humans in content moderation schemes is ethical.

In her dissertation, *Behind the Screen: The Hidden Digital Labor of Commercial Content Moderation*, Sarah Roberts first formally defines commercial content moderation (CCM) as "the organized practice of screening user-generated content (UGC) posted to Internet sites, social, and other online outlets" (S. T. Roberts, 2014). She then reports on the taxonomy or structure of CCM by painting a picture of how and where the work of moderation takes place (S. T. Roberts, 2014). Following the characterization of the industry, Roberts conducts in-depth qualitative interviews with CCM workers to elicit a firsthand account of the various work experiences of these employees. These interviews allow Roberts to humanize the employees and to bring to light their contributions, their sophisticated views, and the conditions under which they labor (S. T. Roberts, 2014). Roberts' comprehensive reporting, research, and exposure of the industry suggests that conditions for commercial content moderation workers should be improved.

In the book, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media,* Tarleton Gillespie, a principal researcher at Microsoft Research and an associate professor at Cornell University, writes an entire chapter dedicated to the human labor of moderation. In this chapter, Gillespie first maps out the structure of human labor in moderation, from the internal moderation teams at the social media platforms to the

users who are encouraged to flag harmful content and all those in between. Specifically, Gillespie reports on the content guidelines Facebook distributes to its moderators who are expected to make decisions using on them. Regarding these guidelines, he writes, “Unlike the clean principles articulated in Facebook’s community standards, they [content guidelines] are a messy and disturbing hodgepodge of parameters, decision trees, and rules of thumb for how to implement those standards in the face of real content” (Gillespie, 2018). After this small case study, Gillespie argues why the task of moderation has traditionally required so many people. He ends the chapter by discussing the labor and logistics of commercial content moderation. Unlike Roberts, Gillespie does not present the challenge of CCM as a personal one with anecdotes, but rather as an institutional one.

Although both of these scholarly works provide a comprehensive review and analysis of commercial content moderation, there is little to no mention of its ethicality. This paper seeks to address this shortcoming in the academic conversation by using a normative ethical framework to make a moral judgement on the use of humans in these roles.

Conceptual Framework

Duty ethics is an ethical framework by which the moral considerations surrounding Facebook’s use of humans as content moderators can be analyzed. This framework will provide a structure through which a moral judgment can be articulately expressed.

Duty ethics is a branch of ethics which judges the morality of an action based on its level of conformity with a moral law, norm or principle. This universally applicable moral law should itself be independent of the consequences of the action in question. Generally, the origin of this moral law can be derived from various sources: secular or religious. Immanuel Kant, an

influential German philosopher during the Age of Enlightenment, formulated a now popular universal moral code known as the categorical imperative.

In establishing an abidance of the categorical imperative, Kant defined a subset of duty ethics known as Kantian ethics. Kant explains the categorical imperative in three different formulations, two of which will be important for the research presented in this paper. The first formulation states the following, "[a]ct only on that maxim which you can at the same time will that it should become a universal law" (Kant, 1785). Essentially this means that an individual should act on some action if he or she foresees a world where everyone is allowed to do such an action. Under this lens, for example, the action of lying would be morally wrong. While a single lie to one person could be argued as inconsequential, if everyone was to lie, our world would cease to be trustworthy. Kant proceeds to identify the second formulation of the categorical imperative which states the following, "[a]ct as to treat humanity, whether in your own person or in that of any other, in every case as an end, never as means only" (Kant, 1785). Like the formulation above, this sentence may be difficult to unpack. Ibo van de Poel and Lamb¿r Royakkers explain this well in their book *Ethics, Technology, and Engineering: An Introduction:*

> In saying they [humans] must never be treated as a means only, he [Kant] means that we must not merely "use" them as means to our selfish ends. They are not objects or instruments to be used. To use people is to disrespect their humanity. (van de Poel & Royakkers, 2011)

Implicit in Kant's categorical imperative is a postulate of equality and an esteem for the autonomy of an individual. For Kant, a violation of this autonomy means an attack on the free-will of the person and thus a disrespect to their humanity.

In the proceeding section, I will use Kantian ethics and both formulations of the categorical imperative to determine if Facebook's use of human employees in commercial content moderation is ethical.

Analysis

Under the ethical framework of duty ethics, Facebook's use of human employees in content moderation schemes violates both formulations of the categorical imperative and thus constitutes the action as unethical and immoral. The following paragraphs will expound on this claim by presenting the specific evidence of Facebook's violation of each formulation.

Violation of the First Formulation of Kant's Categorical Imperative

Through its use of human laborers in its content moderation schemes, Facebook suggests that viewing this harmful content is both simultaneously ethical and unethical. Facebook implicitly determines that it is ethical for moderators to view this content, but unethical for its users. Under the first formulation of Kant's categorical imperative, an action cannot be ethical for some, and unethical for others. Rather, an action that is unethical for one, is unethical for all. Under this framework, Facebook's actions in allowing moderators to view this content while simultaneously barring its users is unethical.

In hiring content moderators or outsourcing the work to a third-party, Facebook takes all the content it wants to shield from its 2.5 billion active users, and it asks a smaller subset of humans to watch it and make decisions regarding it (Facebook, 2020). Is this ethical just because a smaller subset of humans is involved? Facebook would certainly not will that this harmful content be viewed by all users of the platform. In a highly-cited *Wired* article, Adrian Chen writes, "companies like Facebook and Twitter rely on an army of workers employed to

soak up the worst of humanity in order to protect the rest of us" (Chen, 2014). Chen emphasizes that moderators are the subset of humans hired by Facebook to protect the rest of humanity from the harmful and damaging content on social media platforms. The company fails to consider that the same attempted protection granted to "humanity" be extended to the moderators themselves, who are constituents of the same "humanity". In an interview conducted by Isaac Chotiner of *The New Yorker,* Roberts, author of the dissertation cited earlier, also answers on how moderators are the content filters for the rest of mankind:

> What these people were doing was really a front-line decision-making process, where they would sit in front of the screen and jack into a queue system that would serve up to them content that someone else, someone like you or me, might have encountered on the platform and had a problem with. We found it offensive, we found it disturbing, maybe it was really, really bad, or illegal activity, or somebody being harmed. And someone like us would report that. (Chotiner, 2019)

In her answer, Roberts specifically implies that the content reviewed by moderators would be offensive to any arbitrary user on the platform. If it can be offensive to any user, cannot it also be offensive to those reviewing the content?

The first formulation of the categorical imperative necessitates that an action not fit to be universal law for everyone is immoral. The action of Facebook to hire certain individuals to moderate the content that it has deemed dangerous for humanity is to draw lines between those being protected and those who must do the protecting. Facebook implies that while it is immoral that the rest of humanity view this content, moderators themselves are not afforded the same protections. The categorical imperative does not allow an exception of this sort, in fact it is

especially against it. Thus, on the basis of Kant's first formulation of the categorical imperative Facebook's actions are immoral.

In summary, Facebook cannot allow content moderators to view harmful content, while simultaneously disallowing users from viewing it. To do so is to say that viewing the offensive content is both ethical and unethical. Under Kantian ethics, an action that is unethical for even just one person is unethical for all. Despite this, one may argue that Facebook is wholly unaware of the extent of offensive content moderators are exposed to and thus should not be held responsible for immoral action. This argument is flawed on many reasons. First and foremost, it is all together unlikely that Facebook executives and leaders are unaware of the harmful content being posted to its platform. In fact, it is especially because they do know how dangerous the content is that they hire more content moderators. Ellen Silver, Vice President of Operations at Facebook, writes of the increasing magnitude of the operation at the end of 2018, "The teams working on safety and security at Facebook are now over 30,000. About half of this team are content reviewers – a mix of full-time employees, contractors, and companies we partner with" (Silver, 2018). Silver continues to describe the work as "not easy" and points to the distributing and violent content being reviewed (Silver, 2018). There is yet additional evidence that highlights how Big Tech is aware of the problem. Just last month in the *Financial Times*, Madhumita Murgia reports that "[c]ontent moderators working at a European facility for Facebook have been required to sign a form explicitly acknowledging that their job could cause post-traumatic stress disorder" (Murgia, 2020). This article proves Facebook has significant knowledge of the difficulty of content moderation work than the simply "not easy" nature as Silver would suggest. Proven knowledge of the problem heightens the responsibility of Facebook to act morally as they cannot longer hide under a veil of ignorance.

Violation of the Second Formulation of Kant's Categorical Imperative

In employee-based commercial content moderation, workers are used as the means to an end. That end is a Facebook void of all harmful, violent, and disturbing content. To accomplish this task, employees are essentially being used as filters of the platform with little to no respect for their financial needs or psychological well-beings. In this light, Facebook uses content moderators as instruments for its own goal. To do so is to disrespect the humanity of those moderators and is thus, under Kantian ethics, an unethical action.

Facebook uses its workers by not providing sufficient, and adequate psychological services for them; thus, its actions can be judged immoral under the categorical imperative. After consulting with content moderator psychologists, Chen writes that "even with the best counseling, staring into the heart of human darkness exacts a toll. Workers quit because they feel desensitized by the hours of pornography they watch each day and no longer want to be intimate with their spouses" (Chen, 2014). Essentially, because content moderators are exposed to the worst of humanity, it begins affecting their personal lives with family and friends. These workers cease to live their normal lives at home due to what they watch during the day. Unfortunately, it does not stop there. Murgia of the *Financial Times* reports on what employees at these facilities see. She writes that "they had seen multiple instances of severe mental health conditions among their colleagues, and had also been diagnosed with depression themselves, something they believed was exacerbated by their working conditions" (Murgia, 2020). Casey Newton, a journalist for *The Verge*, interviewed dozens of current and former Facebook content moderators at Cognizant in Phoenix. On the workplace conditions, he describes, "It is an environment where workers cope by telling dark jokes about committing suicide, then smoke week during breaks to

numb their emotions" (Newton, 2019). Instead of resorting to positive coping strategies such as counseling and therapy, employees handle the work environment through negative means.

The lack of proper psychological services is even apparent in the statements Facebook executives have put out themselves. Silver writes that "[a]t Facebook we have a team of four clinical psychologists across three regions who are tasked with designing, delivering and evaluating resiliency programs for everyone who works with graphic and objectionable content" (Silver, 2018). The VP of Operations at Facebook seems to suggest that they provide for its employees by providing four psychologists. However, Silver is not clear about how just 4 psychologists will be able to provide for Facebook's content moderation staff of over 15,000 people, nor of how such a task is feasible. Additionally, Silver is not explicit about the 3 regions which the psychologists are to operate in. Are they in proximity to each other or are they spatially distributed around the world? Silver leaves the important answers of these questions to the readers.

Facebook's use of humans for its own selfish ends extends beyond inadequate psychological resources to the compensation package (or lack thereof) offered to content moderation workers. Gillespie writes, "Though some may be troubled by what they're forced to look at, more typically the worker is troubled by whether he can get enough work, whether the pay is enough to support his family, … whether he will be able to afford health insurance this month" (Gillespie, 2018). Sometimes the concern for money by workers supersedes the psychological consequences of the role. Given this, if the role paid sufficient money, perhaps Facebook can make a case that workers are being treated as humans in their own respect. Research via investigative journalism would suggest Facebook does not do this. Newton reports that "[t]he median Facebook employee earns $240,000 annually in salary, bonuses, and stock

options. A content moderator working for Cognizant in Arizona, on the other hand, will earn just $28,800 per year" (Newton, 2019). That amounts to just 12 percent of the median salary of a Facebook employee. Roberts confirms just how dire the situation is for these employees:

> [T]hese were people working at élite Silicon Valley firms. But, instead of coming into those firms as full-badge employees with a career trajectory in front of them, they were coming in through contract labor, third-party sourcing. They were coming in relatively low-wage, especially in relation to any peers that they could be working side-by-side with in such a place. And, in the case of the United States, they didn't have health care provided to them through this arrangement; when we think about psychological issues or other health issues that come up on the job, the way that people get health care is through their employment. (Chotiner, 2019)

To not provide access to healthcare as a job benefit is an egregious negligence on the part of Facebook and its contractors, especially since former employees have and continue to suffer from mental illness. Realizing how low this wage is, Facebook has recently gone on the record stating, "We'll pay at least $22 per hour to all employees of our vendor partners based in the Bay Area, New York City and Washington, D.C.; $20 per hour to those living in Seattle; and $18 per hour in all other metro areas in the US" (Gale, 2019). In Arizona, this would likely entail a salary increase to approximately $37, 000. While this is an increase from the initial $29,000, it is still a fraction of the compensation a traditionally Facebook employee receives.

These salary increases, however, do not extend to those working outside the United States, where the majority of content moderation takes place. "This work [content moderation] is increasingly done in the Philippines…moderators in the Philippines can be hired for a fraction of American wages" (Chen, 2014). Chen then cites an employee in the Philippines being offered

just $312 per month to moderate content for Facebook. Annually, that pay amounts to a salary of just $3,744! In 2014, the same year Chen reports the salary offer for a content moderator, the average Philippines salary was calculated to be at $11,959.56 a year (White, 2014). Even in other countries, Facebook and its contractors are underpaying their content moderators.

The lack of adequate compensation packages and psychological resources given to content moderator workers implicates Facebook as a selfish entity which uses these employees as a means and not as an ends. This is a clear violation of the second formulation of the categorical imperative, and thus Facebook's actions in hiring humans as moderators is immoral. Facebook may defer blame by claiming content moderation workers agree to the conditions of this type of work when they sign up. However, this claim is problematic for a few reasons. Content moderators cannot possibly know apriori what content will appear before them and what psychological conditions, if any they may develop. Content moderators may see one video which is totally innocuous, but the next may be something shocking and abhorrent (Chotiner, 2019). Also, if content moderators did have foresight into the role before accepting, there would not be so many who would leave after just a few years. Roberts says, "The work tended to be something that you would likely not do for a long time… You would either not be able to really take it anymore, or you would become so desensitized as to not be any good at that job anymore" (Chotiner, 2019). All of this is to show that employees are not typically aware of what it is they are to experience at the job. Many are also prompted to accept the role in desperation of an income to make ends meet. It is Facebook's responsibility to ensure a proper care for its employees, so as to respect their autonomy and well-being. Platforms must stand accountable "for making cogent decisions about how this work should be done and by whom, for articulating why moderation should be parceled out in this particular way, and for articulating clearly and

publicly how they plan to make their moderation effective and responsive while also protecting the labor rights and the emotional and social investments of everyone involved" (Gillespie, 2018). To do anything else is to use humans as a means for selfish ends. Roberts puts this selfishness in succinct terms, "the primary function of people doing commercial content moderation at these platforms was for brand management of the social-media platform itself… so that the brand could continue to function as a site where advertisers might want to come" (Chotiner, 2019). This selfishness can be made right if the workers in content moderation are adequately provided for.

Conclusion

I have argued that Facebook's decision to use humans as content moderators of its platform is unethical under Kantian ethics. To the end, I proved Facebook's actions deny not only one but two formulations of Kant's categorical imperative. Facebook's decision to use humans as moderators is not a decision it would will on the rest of humanity. Rather, content moderators are seen as shields for the rest of humanity, as if these workers are somehow less susceptible to the psychological consequences of viewing the material. Kant is explicit that making an exception for a smaller subset of humans to view this content is strictly unethical. Furthermore, to use these workers is to use humans as a means and not as an ends. Facebook's exploit of these workers is evident in its lack to provide adequate psychological services, health insurance, and sufficient financial compensation. Facebook's intention in moderating its platform is a selfish one, as a "cleaner" Facebook potentially draws more advertisers to the site, and thus more profit is generated for Facebook. Facebook's indifference towards its workers even as they

labor for its own selfish ends is the grounds for the violation of the second formulation of the categorical imperative.

In an academic conversation that has lacked ethical analysis pertaining to the use of humans in these roles, this paper presents a small step forward for content moderation workers. Further, this analysis contextualizes just how much dangerous content we would be exposed to if it were not for these people working behind the senses. If anything, the work presented here calls for a change in how social media content moderation operates and in how moderators are treated. While this paper encourages a start to this discussion, additional research is required to feasibly address these calls of action.

References

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

Chotiner, I. (2019, July 5). *The Underworld of Online Content Moderation*. https://www.newyorker.com/news/q-and-a/the-underworld-of-online-content-moderation

Facebook. (2020, January). *Facebook users worldwide 2019*. Statista. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Gale, J. (2019, May 13). An Update on Compensating and Supporting Facebook's Contractors. *About Facebook*. https://about.fb.com/news/2019/05/compensating-and-supporting-contractors/

Gillespie, T. (2018). The human labor of moderation. In *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (pp. 111–140). Yale University Press.

Kant, I. (1785). *The Groundwork of the Metaphysics of Morals*.

Murgia, M. (2020, January 24). *Facebook content moderators required to sign PTSD forms*. Financial Times. https://www.ft.com/content/98aad2f0-3ec9-11ea-a01a-bae547046735

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

Roberts, S. (2019). Understanding Commercial Content Moderation. In *Behind the Screen* (pp. 33–72). Yale University Press; JSTOR. https://www.jstor.org/stable/j.ctvhrcz0v.5

Roberts, S. T. (2014). *Behind the screen: The hidden digital labor of commercial content moderation* [Dissertation]. University of Illinois at Urbana-Champaign.

Rosen, G. (2019, March 20). *A Further Update on New Zealand Terrorist Attack | Facebook Newsroom*. https://newsroom.fb.com/news/2019/03/technical-update-on-new-zealand/

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

van de Poel, I., & Royakkers, L. (2011). *Ethics, Technology, and Engineering: An Introduction*. Wiley-Blackwell.

White, J. (2014, May 15). *What Is The Average Salary In The Philippines?* Dumb Little Man. https://www.dumblittleman.com/what-is-the-average-salary-in-the-philippines/