"We Care About the Internet; We Care About Everything" Understanding Social Media Content Moderators' Mental Models and Support Needs

Sarah T. Roberts
UCLA Department of Gender Studies
sarah.roberts@ucla.edu

Stacy E. Wood
UCLA Center for C2i2
swood@c2i2.ucla.edu

Yvonne Eadon
UNC CITAP
ymeadon@gmail.com

Abstract

Despite the growing prevalence of ML algorithms, NLP, algorithmically-driven content recommender systems and other computational mechanisms on social media platforms, some core and mission-critical functions are nonetheless deeply reliant on the persistence of humans-in-the-loop to both validate computational models in use, and to intervene when those models fail. Perhaps nowhere is this human interaction with/on behalf of computation more key than in social media content moderation, where human capacities for discretion, discernment and the holding of complex mental models of decision-trees and changing policy are called upon hundreds, if not thousands, of times per day. This paper presents findings related to a larger qualitative, interview-based study of an in-house content moderation team (Trust Safety, or TS) at a mid-size, erstwhile niche social platform we call FanBase. Findings indicate that while the FanBase TS team is treated well in terms of support from managers, respect and support from the wider company, and mental health services provided (particularly in comparison to other social media companies), the work of content moderation remains an extremely taxing form of labor that is not adequately compensated or supported.

1. Introduction

Over the past several decades, social media platforms have become central to interpersonal connection, business management, media and news access, entertainment and more. The majority of social media platforms utilize a business model that combines constant, ubiquitous data collection of user activity with algorithmic ranking systems that modulate and personalize content for those same users. The scale and speed of social media platform growth has necessitated the complementary growth of commercial content moderation in order to manage legal liability and community standards while facing increasing pressure to address many of the broader effects of social media.

Commercial content moderation is done at scale by professional, compensated laborers who apply policy set by social media platforms in order to protect companies from public, financial and legal backlash, as well as protect its user-base from a variety of harmful content. This professional community has experienced extremely rapid growth and with it an increased level of visibility and scrutiny. This once ”invisible” labor is now a catalyst for policy intervention, political talking points and immense public pressure related to the way social media platforms do business and the effects of that business on the world.

Commercial content moderation is complex, requiring practitioners to juggle internal policy, community guidelines, legal mandates, a platform's own risk tolerance and the health and welfare of the larger community of users even as these pressures often compete. The work can be traumatic as moderators are inundated daily with violent imagery, child abuse, extreme racism and more. While larger firms often have access to computational tools to aid in partially automating some aspects of moderation, small to midsize firms are often much more heavily reliant on human decision-making to enforce platform rules and norms. Despite advances in machine learning systems that provide companies with ways to automate moderation processes, these processes can still require and often necessitate a human-in-the-loop to make the kinds of nuanced decisions required by company policy, even at firms that can afford to augment their human-based practices with computation.

During the fall of 2020 , our research team conducted semi-structured qualitative interviews with an in-house commercial content moderation team at a mid-size, social media platform we call FanBase. FanBase is made up of millions of chat rooms that range in size from fewer than ten users to hundreds of thousands. We asked nineteen staff about their work practices including questions about: computational and policy tools they used; collaborative and collective knowledge-sharing

URI: https://hdl.handle.net/10125/102883
978-0-9981331-6-4

practices; stamina and well-being; mental models; and external influences on their decision-making.

This paper describes how human commercial content moderators actively contribute to both the design and implementation of policy while navigating the demands of their roles within the context of real-world events. Findings from our research surfaced how dominant outside events were on the working lives of moderators and in particular, a violent white-supremacist event (henceforth ”The Event”) that was planned and coordinated on FanBase. Additionally, FanBase was experiencing exponential and rapid growth during the outset of the COVID-19 global pandemic and as a result team members were taking on expanded roles, contributing to scalable solutions and constantly on-boarding new moderators. Our findings indicate that in the midst of these dual stressors, worker autonomy, peer-to-peer support, financial security and respect within the larger company were fundamental to moderators' mental stamina. As the need for content moderators grows, their work becomes more entangled with computational assistive tools and the breadth of their responsibilities and challenges extends, it is crucial to study not just the factors that go into their decision-making and work practices but also to the environments and mechanisms of support that make their work sustainable and increases quality of life.

2. Background

Content moderation has become a highly contested aspect of social media business practice, having gone from a somewhat behind-the-scenes activity to a central factor in public debates around everything from violent hate speech to disinformation and forcing collective reckoning about the contemporary contours of the public sphere. Research on content moderation has examined a wide-range of phenomena predominately focused on the experiences of the user and policy implications(Caplan et al., 2019; Dori-Hacochen et al., 2012; Milosevic et al., 2018; Pater et al., 2016), but still remains relatively focused on large-scale, US-based platforms. Platforms represent a variety of business models, risk-tolerances, branded corporate identities, sizes, scales, languages, technical sophistication and modes of engagement and so research must address this breadth.

Commercial content moderation involves screening user-generated content to determine whether that content adheres to a platform's policies or not. Moderators deal with text, images and video that can range from identifying and eliminating bots and/or fake accounts, child sexual abuse material (CSAM), hate speech, graphic violence, and/or mis/disinformation. Content moderation is enacted and practiced differently according to the platform and its resources and priorities, but has to an extent remained rooted in policies that reflect people who are predominately white, male, educated and ”technical in skill and worldview.” (Gillespie, 2020) Commercial content moderation necessitates skills beyond rote application of corporate policy and includes ”linguistic and cultural competencies; quick recognition of symbols or latent meanings (Roberts, 2019)” balanced with often competing commitments to perpetual growth and maintaining the integrity of the platform.

Initial attention brought to commercial content moderation exposed what was a less publicly visible aspect of platform brand management, a practice that was often outsourced entirely, underpaid and not well-respected within company culture despite its importance (Arsht, 2018; Gillespie 2020; Grimmelman, 2015; Klonick, 2017; Roberts, 2019). In 2020, Facebook settled a 52 million dollar lawsuit (Dwoskin, 2020) with more than 11,000 employees related to inadequate support and mental health services for commercial content moderation workers who were showing persistent, untreated symptoms of post-traumatic distress disorder (PTSD) stemming from their work on behalf of the platform. Even as this attention to the harmful working conditions of commercial content moderators increased, there has still been little research on the needs of commercial content moderators themselves.

Automated content moderation tools have frequently been posed as the answer to this challenge, either to off-load the more clear-cut cases or to eliminate the need for human content moderation entirely. However, there remain deep questions around how sophisticated automated tools are and will be, whether the problems that come with automation at scale are simply exacerbating the issues at hand, how sustainable these practices are, and what the consequences might be of pushing these decision-making processes into further algorithmic opacity (Gillespie, 8; Marshall, 2021; Myers West, 2018).

In this paper, we foreground the experiences of moderators themselves in order to understand what kinds of support, infrastructure and resources are needed to make this work meaningful and sustainable. The findings and our experiences with the participants also prompted us to think through future work expanding the application of trauma-informed research practices in commercial content moderation research.

3. Methods

We conducted 19 semi-structured qualitative interviews with an in-house moderation team, referred to as the Trust and Safety team, at a mid-sized social platform we pseudonymously call FanBase. This was a grounded theory study (Charmaz, 2006) underpinned by feminist methodology (Marshall, 2021; Reinharz, 1992). Interviews consisted of nine open-ended questions intended to assess moderators' mental models and work environment: computational and policy tools they used; collaborative/collective knowledge-sharing practices; their stamina and well-being; their methods and resources for dealing with the mental and physical impact of their work. We use pseudonyms when referring to individual employees of FanBase. Ten participants had worked for FanBase for seven months or fewer; five had been there between one to two years; and the remaining four had worked at the company between two and a half to three and a half years. Interviews were conducted via Zoom over a week-long period in 2020. After two research team members interviewed participants, interviews were transcribed and coded in Dedoose. Both our team and interview participants recognized a need for a trauma-informed approach to the research process given the nature of their work, which we are working to develop further as an essential intervention into content moderation research in the future.

While participants voluntarily provided some information, our team did not ask directly for interviewees to report their own sexuality, race, disability status, ethnic identity or country of origin. Nonetheless, some of this information emerged during interviews, when moderators themselves invoked aspects of their identity. The team was comprised of U.S. based moderators, leaving linguistic and contextual gaps regarding issues specific to other communities.

4. Findings

Our findings indicate that, at the time of our research, FanBase was attempting to approach content moderation differently from other firms, especially those dominant in the marketplace. The results presented in this section reflect the values, policies, and practices held by the company at one specific point in time, influenced primarily by two separate events: a violent, racist, widely-publicized event with deep ties to the platform ("the Event"), and a rapidly, exponentially expanding user base. This section presents and analyzes the company structure and daily workflows, the profound impact of the Event on FanBase, the dramatic scaling up of the moderation team during the COVID-19 pandemic, the working conditions and mental health of moderators, and areas in which interview participants would like to see improvement both at FanBase and the field of content moderation in general.

4.1. Structure of and Workflow in FanBase Trust Safety

At the time that we conducted our interviews, the Trust and Safety (TS) team was the largest team at FanBase, comprising around forty employees who contended with a user base of around 350 million users and 10 million active users per month, according to TS Specialist Connell Murphy. In the FanBase model, idividual chat rooms also rely on community-driven moderation, with differential standards set by the chat room community members themselves, augmenting and complementing Trust and Safety's work.

TS positions at FanBase are organized according to experience. Analysts are the newest and least experienced employees. After 6 months, analysts become Specialists, choose a specialization, and begin an individual or group research project concerning their chosen area of specialization. Senior Specialists, Team Leads and Managers all have more responsibility and autonomy, and at times elect to form Specialist teams or "squads," which include the policy team and content specific teams including: the anti-extremism squad, the child sexual abuse material/child sexual exploitation (CSAM or CSE) squad, and the cybercrime, misinformation, spam, gore, non-consensual pornography (NCP), and self-harm squads. The work involves both the proactive work of keeping an eye on potentially volatile chat rooms, and the reactive work of evaluating user-submitted tickets. Tickets come through queues that are specific to certain types of content, labeled by users themselves.

Onboarding is a sixty-day process introducing the team's operational and philosophical approach to the work including classes, shadowing, reverse shadowing, and support from trainers.

Established moderators are expected to "action" (i.e., respond to) 500 tickets per week. When reading the ticket report, Anti-Extremism Team Lead Katie Warner uses a mental triaging process to try to

> ...understand as quickly as possible, like what it is that this person is reporting? Are they reporting a user? Are they reporting a message? Are they reporting an image? Are they reporting an entire chat room? And then depending on what they are reporting and why they are reporting it,

> that then frames for me what it is that I'm looking for, to determine whether there is a violation or not.

Katie's evaluation process always starts with a set of mental steps and actions she undertakes with evaluating a ticket. Her first evaluation is to assess the type of content being reported, as that shapes how she approaches the next steps of evaluating whether or not the content has violated FanBase's terms of service. This initial mental triaging is important, as it determines which pathway to take through a mental map of evaluative steps, questions posed and policies invoked for each different case and violation type. Importantly, moderators are able to opt into or out of different types of content. This has significance not just for workflow or decision-making but for managing exposure to traumatic content and stamina. For Analyst Bailey Bishop,

> It depends on what I'm tackling, the tools I would use to handle, for example, a regular spam ticket. If a user is spamming it's going to be vastly different from the tools that I use to, let's say, evaluate whether this user is like sending a malicious link that's intending to hack other users or whether this user is an extremist that has to be handled.

Katie also described weighing the privacy of individual users against the likelihood of a given space containing high-harm hate speech. Extremism tickets are thus often much more involved than other types of tickets.

Other participants also described their workflows in terms of the questions they asked themselves in the process. Bailey Bishop recounted her workflow as she responded to a spam ticket: she first opens the ticket, then looks at the message links, and then verifies that the content is what the reporter says it is. To do this, she goes into ”admin mode,” a technical layer that provides context and metadata around the reported messages. After verifying the ticket, she uses a virus checking tool to see whether or not the link was indeed malicious. If the link was malicious, she bans the user and deletes the link. Katie also mentioned the importance of comparing the report with the information visible in admin mode in order to verify the ticket.

Analyst Miles Pham discussed the process around evaluating gore tickets: ”...the great and bad thing about gore is that like, you kind of know very quickly in like two seconds, not even two seconds, that, Oh, God, even just feeling that visceral reaction. It's like, Yeah, that's good enough to, to either warn, or delete, or ban this user...” Noting the viscerality of seeing gore, Miles' statement illustrates the marked difference between severe content, which can have significant mental impact on moderators, and something that is much lower harm and less stressful to deal with, like spam.

Legal counsel and Head of Policy Jordan Davis used a mental model based on a judicial system metaphor to explain the way he approaches FanBase's content moderation policies and practices:

> My initial foundation of the team was that we should all be federal judges... We should understand that we are there to resolve each case in front of us, but we are also there to think about how... each policy affects society at large... and strike down laws if they are unconstitutional, so to speak...The goal is to have autonomy radiate outwards.

This metaphor demonstrates that moderators themselves assign gravity to their work and are aware of the power they hold, as well as acknowledging the iterative and deliberative nature of their work. Importantly, it demonstrates that understanding pre-existing frameworks moderators bring to their work is essential to making sense of their mental triaging, logic and decision-making. We find this to be an area that requires more interrogation across all the teams in a firm that create and execute content policy.

Bailey Bishop and other analysts discussed the collaborative nature of the TS team, and the fact that Analysts and Specialists can always ask one another for a second opinion or attend their managers' office hours. In Bailey's words:

> ...part of the reason we have a specific Trust and Safety chat room is because you know, if I have a question about a certain ticket, or I need someone with more experience, I can just ask the room...So you can just go in and ask any questions about any tickets that you have...what you're going to find if you work here is that a lot of times, it's just really about asking questions. And it's very normalized to like, ask and get a second opinion in the team. To the point that like, for example, if you're like, before you delete any large rooms, you go to the what's we have a second opinion channel, and make sure you get a plus one before you delete any very large room just to make sure.

Asking questions, of oneself and of others, is an integral part of the job, especially for newer hires.

Collaborative evaluation of content is encouraged, especially in particularly difficult cases. Analyst Adrian Williams summarized this kind of circumstance:

> I'm thinking of a case we tackled as a group recently... it was a discussion over the sort of misinformation going out around the Stop the Steal for the election, where we were seeing Facebook and Twitter start really bringing down the banhammer a little too late on some of these misinformation groups. And we were starting to see a bit of an uptick of it on our chat rooms that are right wing spaces, and the debate was: should we action individuals for spreading this fake information around? Should we action rooms for allowing this information to be spread? How do we communicate that, and at what severity, if any, do we take action?... The big debate was over like most of us, as far as I can tell, all of us are quite left-leaning. And so my point was, like, we are overcompensating for perceived anti-right-wing bias when interacting with these rooms... So we had to consider like, A) company image. B) are we actually enforcing our policies fairly on these spaces? Are we giving them too much leeway? Or is our kind of left-wing leaning making this more emotional for us than maybe it should be? And then, what do we actually do about this information? And that debate lasted for an hour or two hours. And I don't think a real conclusion was actually reached. But I think we ended up just warning some of the individual users: look, this is fake, don't spread it around.

In this case described by Adrian, the attempt to come to a team-wide consensus on the issue was not accomplished, but rather evolved into a first attempt at creating a company-wide policy for how to evaluate misinformation as a type of harmful content. It involved the people on the ground-those who actually moderate content day-to-day, no matter how long they had been at the company or level of experience. It could have been a slightly easier discussion among a smaller group of higher-ups, but in this case, FanBase valued the expertise and experience of their front-line moderators.

Several interview participants spoke about how they felt respected by FanBase as a company and compared the company's treatment of content moderators to how other tech companies. From Supervisor Tate Bronson's perspective there is "a lot of support for what our team does. And at other tech companies, content moderators are literally put in the basement. Like I remember at my last tech company, the team that looked at CSAM, they had to really fight not to be put in the basement." Other participants echoed this, pointing out that FanBase invests a lot of time and money into their team. For Specialist Amir Bahar, "...what I think is wonderful about working in Trust and Safety at FanBase is that FanBase doesn't make the rules, Trust and Safety at FanBase makes the rules. And that is incredibly valuable." Not only are individual content moderators given a lot of agency at the company, but the team as a whole is allowed adequate autonomy and mental space to manage itself. For Roy B., the company makes Trust Safety feel it is "not an afterthought, we are integral and vital to what's going on."

Policy Team Lead Jordan Davis described the work of the TS as being predicated on "the affordances and constraints of the actual FanBase platform and service." Some chat rooms are very active, and others are quieter, meaning some need more attention from the TS team than others. Loren Smith, a member of the policy team, considered size to only be part of the equation: "I wonder, too, if there's something kind of structural about it, that allows for the ability for your team to be more strategic about intervention, in a way that like. . because it's not about size, but it's actually about the way that communities are configured that might allow you to kind of grasp context quicker or, or understand relationships faster, in a way that might not be as possible on Twitter." For Loren, the chat-room-based structure of the company made evaluating context, including assessing relationships between users, easier for content moderators when compared to larger social media platforms.

Because FanBase consists of millions of individual chat rooms, the nature of a specific chat room may be more easily grasped than more amorphous communities that form on a hyper-visible individualized platform like Twitter, or a hybrid arena like Facebook. Jordan Davis theorized that part of this has to do with being a so-called "third wave" Internet platform. The first wave, which included the current giants like Twitter and Facebook, asked "What if everyone could talk to everyone?" The second wave, including Pinterest and Snapchat, considered alternatives to that model. FanBase, as part of the third wave, considers the question "What if we made it about communication instead of just about content?" Jordan concluded this theorizing by stating that, as Fanbase grew and became well known, the focus on communication over content "really informed our thinking both on the product side and the policy side of what we do today."

4.2. The Event

A few years prior to when we conducted our interviews in late 2020 , far-right extremists used FanBase chat rooms to plan and execute a violent white supremacist incident. That the Event was planned on FanBase was also widely publicized and had immediate reputational repercussions; the platform had first been niche, known for its role in fan and gaming culture, but after the Event, its reputation became associated with right-wing extremism. Jordan considered the Event to be ”... a crystallizing moment for Trust and Safety and for FanBase as a whole.” Katie Warner, the team leader of the anti-extremism team, discussed the catalyzing violent national event’s emotional and organizational impact on TS employees:

> The narrative of the Event and what happened with FanBase’s role in [it] still, I think, weighs very heavily on everyone, it’s like narrative that is mentioned again, and again, it comes up in the media all the time–that’s really how the [anti-extremism] team was founded, it was to prevent another Event. And that has kind of been the focus of the team.

The anti-extremism team was founded as a result of the Event having been planned on FanBase, and the reverberations of it still affect the team to this day.

Amir Bahar echoed this and connected it to the larger brand management of the company ”...if we are a little bit over-prescriptive in our action, it’s better than being under-prescriptive, and someone dying, and then FanBase is in the newspaper.” Given that the repercussions of the violent incident led to restructuring of moderation teams, the Event shaped, at an organizational level, policy decisions and a collective orientation to become a less permissive environment for extremist activity on the platform. It also influenced the actions of individual FanBase moderators for whom the Event cast a persistent shadow, informing how they mentally evaluated outcomes of action versus inaction with regard to moderation discussions.

As a result of the Event and the formation of the anti-extremism team, FanBase went through a first wave of scaling up its moderation workforce. Tate even described wanting to work for FanBase because of the Event: ”I knew about the Nazis that had used FanBase to plan the Event. And that was a big part of why I wanted to come to FanBase. I am Jewish, and I don’t like Nazis. I don’t like white supremacy.” Tate felt called to work against the kind of online hate groups that led to organized mass physical violence, in particular because of their own identity as a Jewish person.

Tate described their mental framing of moderation work as a calling, giving another example of misuse of the platform that they were keen to eradicate.

> So when I first started, I think that a lot of early employees who were there [when the Event happened] were there for the first time that journalists started reporting that there was [child sexual abuse material] on FanBase. I think people were really impacted by that very early on. Because, you know, as a small startup, where so many people felt like they put a lot of themselves into it. And it felt, I think, painful to people to discover that it was being used in those ways.

The negative attention that came to FanBase as a result of the Event revealed other types of content that was being circulated on the platform. The small size of the company, beyond being a factor in the planning of the Event (in that it was able to slip through the cracks), also influenced employees’ affective relationship toward FanBase itself: not only did FanBase employees not want to be associated with a violent extremist incident, but they also did not want their company and its product to be associated with it. At every level of the company, up to and including the CEO, there seemed to be genuine remorse and a sense of responsibility that the Event had occurred and was so closely linked with FanBase.

Other team members discussed how pervasive the Event was in motivating the work that they do. Tom Morland, Senior Trust Safety Specialist, said:

> We want to prevent another Event; we want to prevent, you know, any type of violence that we can. And we do feel like primarily, they’ve gone to other platforms... But those people still hang out on FanBase. So you know, we’ve got to be vigilant.

Amir Bahar described similar feelings of responsibility shared by many moderators for the overall tenor and reputation of FanBase; in addition to the psychologically challenging material they had to view, this sentiment increased workplace pressure.

> I think we have full autonomy here to say that we don’t want [extremism] to be organized here. Especially since bad things can happen. And we don’t want bad things to happen. And then people say, well, they did it on FanBase first. I think it’s like, it’s our responsibility to not be in the news for something like this.

4.3. Scaling up during the COVID-19 pandemic

Over the course of the COVID-19 pandemic, FanBase experienced an explosion in their user base, resulting in a second wave of TS hiring. The TS team hired sixteen new analysts during the pandemic, all of whom were only six months into their tenure at the time of the interviews. In Tate's words, "Because of the rate at which things just skyrocketed during COVID, we simply can't train people fast enough to meet the demand." Tate went on to discuss the challenges of hiring so many people in such a short time, which included the difficulties of creating a sustainable management infrastructure. They nevertheless expressed amazement at the overall success of the endeavor. Amir Bahar calls the expansion of the team "a remarkable feat of human capacity." The expansion put more of a burden on people who were training the new hires, like Katie Warner. After working "crazy amounts of overtime," Katie looked back on the hiring explosion: "But it's almost like it's just been a big black box of like time between, 'Oh, yeah, we were a team of 23. And now suddenly, we're a team of 40 ...it was a lot of work, and a lot of struggle."

TS Analyst James Armstrong described the paradox of the user growth under COVID:

> I would say that, you know, since I started six months ago, FanBase has grown a lot... by about 20 million [users in a short period of time]...So like, how does a company that's a startup company with less than 250 employees-How does it manage like a user base increase like that, like everyone at the company is always saying, 'Oh, we want to be Facebook, we want to have a billion users.' But if we had a billion users right now, we would need to triple [our team]. It's something that I think about a lot, and I kind of worry about a lot.

Yet not every employee bore the influx of users and growth in the same way. Each participant had a unique experience of the scaling up that was happening as a result of the pandemic; some were mentally burdened by it, reporting increased workload and increased stress; while others barely noticed it was happening. For the newest hires, the explosion in growth of FanBase was the status quo, and they had no prior, smaller work environment and experience with which to compare their current one.

Other aspects of work life were implicated in the growth, beyond an increase in moderation tickets. Amir noted that team-wide communication became more difficult as the team grew:

> When I joined, there was only I think I was like, number thirteen or fourteen on Trust and Safety. And we didn't really need to have a whole lot of formality because like, 10 of us sit in a circle, and we talk about something for like, 10, 20 minutes, and we all have an understanding, we can just walk away, no one has to write anything down. We don't have to make a formal policy change, we don't really have to make an announcement, because everyone was physically there. It's a little bit different when you've got 40 people on the team...It's been a challenge for some because communication sort of falls by the wayside a little bit."

As Amir implied, the pandemic expansion resulted in some significant organizational restructuring, including the formation of a formal policy team as a means of connecting to larger conversations about moderation outside of the company and to formalize internal processes. Scaling up significantly resulted in a need for standardization and coordination across the moderation team; the casual discussions that Amir remembered taking place began to give way to uniform decision-making in the face of a ballooning user base and a TS team that was growing in parallel.

Although TS had to contend with some users accusing them of being too quick to moderate, the risk of another Event happening was enough for the team to err on the side of taking action. Moderators had to constantly balance their commitment to safety for the majority of users, with complaints and demands about the decisions they made, their own instincts, and company policy. Most moderators demonstrated a serious commitment toward their vocation; echoing Tate, it seemed to approach a moralized sort of calling. The psychological responsibility to make good decisions while protecting the userbase and, perhaps, the wider world, became the purview of the FanBase moderators, expanding their mental triaging and gaming out of consequences well beyond the bounds of earlier days.

4.4. Labor Practices, Working Conditions and Their Impact on Moderators

Two major topics arose in interviews around working conditions and employer support: financial compensation and mental health support. While it would be overstating the case to say that our findings allow us to generalize across the industry, the majority of our participants had worked in multiple Trust and Safety positions and compared their experiences at FanBase within that context, providing a vector for such comparisons to be made.

Content moderation is not a field typically known for generous pay or benefits, or even job stability (Klonick, 2017). FanBase is relatively unusual in its use of an in-house moderation team and higher-than-industry-average wages. Jordan Davis, legal counsel and head of the policy team, says, ”The average salary that we paid Trust and Safety here is $90,000 a year... This is a skilled labor set.”

The recognition of content moderation as skilled labor, rather than an afterthought to be outsourced or contracted away for low pay and low prestige, is a departure from the framing of the work in many other firms. Yet, not only does content moderation require a certain level of technical knowledge and expertise, but it also requires strong analytical skills in applying policy and using evidence for decision-making, in addition to emotional and mental skills like empathy, and the ability to cope with frequent and prolonged exposure to traumatic content.

Loren Smith, a policy team member and former Trust and Safety analyst, describes their experience with graduate school hierarchy between the humanities and STEM students and then relates that division to a similar one they see happening at FanBase and other companies:

> I see that same divide between content moderation, and engineers slash product managers. It’s like, we can have shadow sessions with our [engineering] team as much as we want. But they do not do [content moderation] day in day out; they don’t understand the issues in the same way that we do.

A number of participants acknowledged that while their compensation was, relative to the rest of their field, generous, wages were still lower than those of the company’s engineers, and this discrepancy could be upsetting. From Katie Warner:

> Does the company leadership truly actually recognize the value that we are bringing and like what would happen if we all disappeared, and to think of how much all of the engineers are being paid compared to what we are all being paid? It’s a very hard pill to swallow and one that I often can’t dwell on too much because I don’t think I would be–I don’t think I could physically, mentally, emotionally, physically, like get myself through work every day if I thought too much about it.

Analyst Miles Pham brought up the need for mental healthcare for content moderators that focuses on PTSD and dealing with trauma stemming from their day-to-day work.

> I personally would love more resources; like, I’ve independently on my own time looked for resources around dealing with PTSD and trauma stress related incidents...Like, I’ve looked up Peter Levine’s book *Waking the Tiger* that focuses a lot on this trauma, and it’s helped me a lot. I do with is was more accessible and like, more intentionally put into Trust and Safety Teams. I do think we already kind of do that with our wellness and sprinkling it [in], but it almost feels like its just enough [to] further trigger people, but it’s not enough to really provide people with the resources and skills to address it head on...But I wonder if I’m more of an edge case, because I find myself gravitating to this higher harm - like CSAM - work...

Other participants mentioned that they or other team members see therapists outside of what is available at FanBase, so there may have been employees already receiving trauma-informed mental healthcare, but it was not a formalized component of the resources available through the company. Tom Morland reflected on the relationship between his mental health, content moderation duties and financial wellness:

> But like [getting paid more than content moderators at other companies] doesn’t change the reality of our situation. So yeah, I mean, pay us enough to, you know, make a living for ourselves here because for me, like, the trauma I experienced at my [previous employer] was severe. Like I mentioned, I can’t even watch, you know, certain types of violent content in PG-13 movies against women anymore. It’s very severe. But you know, I feel in some ways almost like a soldier returning from war kind of, like I’ve put in six years of my life into really traumatic work, but I don’t have much to show for it financially. And that’s difficult.

Several participants also described financial well-being and mental health as going hand-in-hand. When FanBase doubled the size of its team after the onset of COVID-19, most team members were paid hourly rather than salaried. Some participants had initial misgivings about this arrangement and what it might signify. Analyst Stephanie Bishop reflected on it this way:

> ...nobody else in the company really understands the type of work that we do because they were expecting us to keep track of hours, don't do any overtime, work 40 hours a week and then go home. And it's you know....we realize that's not something that we can necessarily do. There's escalations, there's emergencies, there's all these different things where it's not like, I'm an engineer, and I can just turn off a laptop at 5pm to be like, "Alright, see you guys tomorrow, I'm done."

This frustration with the nature of the work and the desire to achieve a means of compensation that allows for work-life balance came up repeatedly, as did the availability of professional mental healthcare provided by FanBase. Moderators also engaged personal strategies for mental well-being, such as compartmentalizing. Analyst Baily Bishop was happy that FanBase seemed to put "a huge emphasis on mental health support," including offering unlimited sick days for employees, but it was clear that overall mental health and wellness was an ongoing issue for all the moderators we spoke to, and was a higher-order concern that could certainly impact moderators' mental stamina, efficacy and resilience when it went un- or underaddressed.

5. Conclusion

The 19 interviews and many hours spent in conversation with FanBase's content moderation team yielded both expected and unexpected results related to the ways in which human moderators actively and reflexively contribute to the design and implementation of company moderation policy and practice. We witnessed the FanBase Trust and Safety team struggling to keep up with itrs explosive growth amidst the ongoing COVID-19 global pandemic, an external and unforeseen global crisis. But they struggled, too, with the lasting psychological traces of internally-grown crises, such as the platform's centrality to the organizing of the far-right violent Event that resulted in injury and death. In the shadow of that incident, employees disclosed a deep sense of responsibility that framed and informed their moderation work, not only with respect to preventing another Event but also in keeping the platform a generally hospitable place for its users. This was in addition to the everyday requirements of social media moderation: removing some of the most abusive content the internet has on offer. Yet, paradoxically, this mission-critical practice went undervalued, even at FanBase: although moderators' base pay exceeded the standard of Trust and Safety moderation roles at other platforms, salaries were notably lower than those of other staff members, creating hierarchies that privilege technical skills over the complex mental modeling, decision-making, analytical skill, emotional intelligence, contextual intelligence and flexibility required in moderation work.

Ultimately, this phenomenon contributes to commercial content moderation's undervaluation, despite its centrality to keeping the platform and, by extension, society safe. FanBase's moderation team demonstrated sophistication, empathy and care in the decisions they made, and took seriously the material with which they engage. This research demonstrates, however, that even at a company that puts relative high value on its content moderation practices (via higher than normal compensation, keeping the work in-house and full-time), there is a gap yet to be filled in internal organizational and public understanding of content moderation and its role in the social media ecosystem, as well as the impact it has on those who undertake it. With that understanding we hope to see greater value, measured in remuneration, support and social recognition, for the work content moderators do on all of our behalf.

6. References

Arsht, A., Etcovitch, D. (2018). The human cost of online content moderation. *Harvard Journal of Law and Technology*.

Caplan, R. (2019). Content or context moderation? Artisanal, community-reliant, and industrial approaches [Report]. Data & Society.

Caplan, R., Hanson, L., and Donovan, J. (2018) Dead reckoning: Navigating content moderation after "fake news."

Charmaz, K. (2006). Constructing Grounded Theory. SAGE, Thousand Oaks, CA.

DeVault, M. Gross, G. (2012). Feminist qualitative interviewing: experience, talk, and knowledge. *Handbook of feminist research: Theory and praxis* SAGE.

Dori-Hacohen, S., Sung, K., Chou, J., Lustig-Gonzalez, J. (2021). Restoring healthy online discourse by detecting and reducing controversy, misinformation, and toxicity online. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Dwoskin, E. (2020). Facebook will pay millions to content moderators who suffer PTSD at work. *The Washington Post*.

Gillespie, T. (2018) Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

Gillespie, T. (2020). Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2).

Grimmelman, J. (2015). The virtues of moderation. *Yale JL & Tech*.

Hesse-Biber, S. N., Leavy, P., Yaiser, M. L. (2004). Feminist approaches to research as a process: Reconceptualizing epistemology, methodology, and method. *Feminist perspectives on social research*, 41, 3-26.

Klonick, K. (2017). The new governers: The people, rules and processes governing online speech. *Harv. L. Rev*.

Marshall, B. (2021) ”Algorithmic misogynoir in content moderation practice.” [Report] Heinrich-Boll-Stiftung European Union.

Milosevic, T. (2018) Protecting Children Online?: Cyberbullying Policies of Social Media Companies. The MIT Press.

Myers West, S. (2018) ”User interpretations of content moderation on social media platforms.” *New Media & Society*.

Pater, J.A., Moon, K.K., Mynatt, E.D. and Fiesler, C. Characterizations of online harassment: Comparing policies across social media platforms. *Proceedings of the 19th International Conference on Supporting Group Work*.

Reinharz, S.(1992). Feminist methods in social research. Oxford University Press.

Roberts, S. T. (2019). Behind the screen: Content moderation in the shadows of social media. Yale University Press.