Article

new media & society
1–17
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/14614448221122224
journals.sagepub.com/home/nms
SAGE

"I'm not this Person": Racism, content moderators, and protecting and denying voice online

Rae Jereza
New Jersey Institute of Technology, USA

Abstract
Much scholarship across the humanities and social sciences seek to shed light on the intersection of far-right politics and social media platforms. Yet, scholars tend to focus on racist actors and the ideological underpinnings of platform policies while the contingencies that shape the experiences of content reviewers who make decisions about racist content remain underexamined. This article fills this gap by exploring such contingencies from a linguistic anthropological perspective. Drawing on Facebook moderators' stories, I illustrate the factors adjacent to, and beyond, ideology that animate the adjudication of racist hate speech.

Keywords
Affect, commercial content moderation, Facebook, far-right, platform governance, stance

Introduction

One afternoon in April 2021, I found myself on a radio show defending Facebook content policy and the existence of systemic racism against a white radio host, who asserted that she was "censored" on Instagram for questioning the efficacy of masks in protecting against Covid-19. She reiterated a contemporary right-wing refrain that Facebook constrained her "freedom of speech" and that people had a right to access "both sides" of every story. Unaccustomed to speaking on radio shows and caught off guard by

Corresponding author:
Rae Jereza, Department of Informatics, New Jersey Institute of Technology, GITC, Room 3500, University Heights, Newark, NJ 07102, USA.
Email: jereza@american.edu

her right-wing stance, I clumsily responded by pointing out that freedom of speech is unevenly protected, and that people of color have been incarcerated and killed for exercising their "freedom of speech." In other words, freedom of speech is not politically neutral and is shaped by systemic racism.

"But my Black friends say systemic racism doesn't exist," she replied.

"I think you ought to read the vast scholarship on systemic racism," I retorted. "And also," I continued, "if anything, Facebook is too lenient. It is in their business interest to allow as much speech as possible from across the mainstream political spectrum. They are not targeting you."

The other guest, a former Facebook moderator and vocal critic of the social media giant, emphasized the unlikelihood of her suggestion that Facebook deliberately censored her account. As he noted, moderators, precarious workers who review most user-generated content, are constrained by productivity metrics. Moderators thus do not have the ability to single out certain accounts, nor do they have the power, time, or energy to systematically rid the platform of content from certain political perspectives.

I highlight this interaction because it captures so much of the conflicting, contemporary discourse regarding political content, racism, and platform governance (Gillespie, 2017) in the United States. In this dialogue, the white radio host uses a right-wing defense of online expression abstracted from a broader context of power relations and racial violence. Meanwhile, the queer, Filipino, socialist, anthropologist is frustrated by this erasure and calls her out instead of engaging directly with her claim. The content moderator, whose labor is obscured in contemporary discussions of platform "censorship," points out that those who adjudicate objectionable content for low pay have little reason or capability to "censor." Instead, those workers are preoccupied by following Facebook policy "accurately" to pass routine audits on which their job security relies.

As of May 2021, journalists report that around 15,000 people moderate Facebook and Instagram around the world (Messenger and Simmons, 2021). Most commercial moderators are outsourced, offshore workers in the Global South, who are precariously employed, underpaid, and endure difficult working conditions. Social media corporations rely on moderators' labor to keep "objectionable" content like hate speech off their platforms, thus keeping platforms usable and sustaining companies' ads-targeting business model (Gillespie, 2017).

Much scholarship across the humanities and social sciences seek to shed light on the intersection of far-right politics and social media platforms. Yet, scholars tend to focus on racist actors and the ideological underpinnings of platform policies while the contingencies that shape the experiences of content reviewers who make decisions about racist content remain underexamined.[1] This article fills this gap by exploring such contingencies from a linguistic anthropological perspective. Drawing on Facebook moderators' stories, I illustrate the factors adjacent to, and beyond, ideology that animate the adjudication of racist hate speech.

These stories come from a larger ethnographic project on Facebook content moderation as practiced in four different third-party vendor sites in the United States and Ireland, and as understood by Western policy workers, journalists, and non-profits from 2020 to 2021. This article draws on semi-structured interviews, email correspondences, messages, and phone calls with 6 of the 30 Facebook moderators with whom I corresponded between 2020 and 2021. From our correspondences, I gleaned the interactional, discursive, and material processes that animate moderators' decision-making. Out of all the moderators I got to know during this ethnography, only 6 dealt with hate speech on a regular basis, while the latter 24 worked almost exclusively in other subject-matter areas.

As I show in my analysis, interlocutors managed fears of potential job loss through strategically, and often ambivalently, aligning with political stances they found flawed or limited. Such alignments caused them distress and frequent encounters with racist hate speech made them doubt their own and their coworkers' politics. By highlighting the politically fraught dimensions of moderators' work, this article extends previous work on moderators' well-being (Chen, 2014; Breslow, 2018; Jereza, 2021; Newton, 2019; Roberts, 2016, 2018, 2019).

Moderators' stories suggest that adjudicating content does not necessarily flow neatly from ideological stance to decision. Rather, moderators act as human conduits of "ordinary affects" (Stewart, 2007) that animate the ways in which racism is understood and operationalized. Moderators' bodies are the sites through which contemporary anxieties, power relations, and contradictory ideological investments surrounding race and racism in the United States pass through and are ambivalently translated. In Kathleen Stewart's (2007) terms, my interlocutors feel and carry "an animate circuit . . . where the overdeterminations of circulations, events, conditions, technologies, and flows of power literally take place" (p. 3). Put simply, what gets classified as hateful, racist content, and removed from the platform is conditioned by Facebook's neoliberal business model as much as it is shaped by post-racialism:

> . . . a twenty-first-century ideology that reflects a belief that due to the significant racial progress that has been made, the state need not engage in race-based decision-making or adopt race-based remedies, and that civil society should eschew race as a central organizing principle of social action. (Cho, 2009: 1594)

An anthropological perspective offers us a way out of the binaries, for example, censorship/amplification, that plague discourses surrounding online speech and politics, evident in the radio show debate with which I began this article. While high-level Facebook executives do make unilateral decisions that privilege historically dominant groups,[2] the daily adjudication of hate speech is primarily a task for commercial content moderators who negotiate their own political beliefs, corporate policies, the ever-present threat of job loss, the uncertainties of an increasingly visible white supremacist movement, and the visibility of anti-racist movements through "stancetaking," which linguistic anthropologist Alexandra Jaffe (2009) defines as "taking up a position with respect to the form or the content of one's utterance" in ways that "reproduce (or challenge) social, political, and moral hierarchies in different cultural contexts" (p. 3). This concept allows us to capture the dynamism of content moderators' decision-making processes.

This article thus also demonstrates the usefulness of linguistic anthropological tools in analyzing how moderators encounter and evaluate hateful posts. Through examining their narrated experiences, I show how working conditions can lead moderators to become complicit in reproducing meanings of racism abstracted from the history, and ongoing reality, of white supremacy in the United States. Then, I demonstrate how the structure of content moderation, as practiced by Facebook's third-party vendor sites, constitutes fertile ground to produce troubling doubts about coworkers' stances (Du Bois, 2007; Jaffe, 2009) as potential racists. I conclude by reflecting on what moderators' interactions with—and interpretations of—racism and racist content suggest about content moderation as a site of contestation between far-right groups online and mainstream liberal society.

Reviewing hateful content on commercially moderated platforms

Sarah T. Roberts (2019) describes commercial content moderators as professionals "paid to screen content uploaded to the internet's social media sites on behalf of the firms that solicit participation" (p. 1). Although this article focuses on moderating hate speech, it is important to note that moderators review many different types of content, ranging from the relatively mundane (e.g. spam and intellectual property violations) to the disturbing and graphic (e.g. sexual abuse, physical violence, bullying, harassment, and more). Many suffer from emotional and psychological distress due to viewing graphic content without adequate mental health support combined with productivity pressures (Chen, 2014; Breslow, 2018; Jereza, 2021; Newton, 2019; Roberts, 2016, 2018, 2019). Because moderators operationalize platform policies, they play a crucial role in platform governance: "the layers of governance relationships structuring interactions between key parties in today's platform society, including platform companies, users, advertisers, governments, and other political actors" (Gorwa, 2019: 854). They keep online worlds functioning by acting as "silent processors" (Carmi, 2019) and "custodians" (Gillespie, 2018) who perform the crucial task of brand management (Roberts, 2019).

Yet relatively little is known about *how* commercial moderators review what platforms often refer to as hate speech. Roberts' (2019) book on commercial content moderation, *Behind the Screen*, is an important exception in that it offers insight into the process. In the book, she describes how moderators must decide "between their own values and those of their client, and to be able to compartmentalize the former while on the job, in favor of working from the perspective of the latter" (Roberts, 2019: 145). Her interviewee, Rick, notes that it can be challenging to "put aside your personal philosophy, your beliefs, your creed, and moderating to the client's wishes" when reviewing posts relating to politically contested topics like gun control or racist and homophobic posts (p. 145). This article builds on Roberts' work by focusing on what it can look like to "put aside" one's beliefs. As I illustrate, "belief" involves negotiating politically charged affects.

Scholarly analyses of policy documents, platform design, algorithmic tools, and other technological affordances also offer insight into how platform governance mechanisms shape how platforms treat hateful, especially racist, content. Ariadna Matamoros-Fernandez (2017) and Eugenia Siapera and Paloma Viejo-Otero (2021) demonstrate how platform policies, infrastructure, and software design ignore power differentials between different races, genders, and sexualities and frame forms of social inequality as categories of difference. As Siapera and Viejo-Otera (2021) argue of Facebook's Community

Guidelines, this approach "repeats the same unfair treatment to which racialized people have been subjected" (p. 127), aligning with the broader post-racial turn in American race-thinking (Bonilla-Silva, 2015). Bharatha Ganesh (2021) furthers this critique, using a concept called "platform racism" to contend that "by adopting the most minimal definition of white supremacy in its design of policies . . . Facebook's software and infrastructure enabled the persistence of white supremacists on its platform" (p. 2). Here, Ganesh is referring to the findings of the most recent Facebook's Civil Rights Audit Report,[3] in which Civil Rights auditors criticize Facebook for prohibiting only "content expressly using the phrase(s) 'white nationalism' or 'white separatism.'" Concerned with operability and profits over combatting racism, Facebook effectively reproduces what Eduardo Bonilla-Silva (2010) terms "racism without racists" or "the practices and mechanisms that reproduce racial inequality and white privilege" (cf. Ganesh, 2021: 2).

Such scholarship reveals how post-racial assumptions and profit motives animate platform policies, infrastructure, and software design. Yet, they have typically relied on analyses of publicly available policy documents and statements from company executives. There is therefore less known about how lower-paid employees at the front lines of platform governance negotiate ideologies that appear coherent and consistent in public-facing documents and statements. The ways in which employees' own economic concerns interact with platforms' corporate goals are also unclear. I argue that pursuing an analysis grounded in the contingencies of content moderators' experiences allows us to explore the spaces in between ideology and profit, on the one hand, and the realities of operationalization, on the other, to reveal the contradictory investments, concerns, and organizational factors informing the adjudication of hateful content not captured by hate speech policy. In other words, I show how moderators are made complicit in platform and platformed racism through productivity metrics that discipline them into aligning with post-racialism.

Navigating affects through stancetaking

In *Cruel Optimism*, Berlant (2011) uses affect to explore collective atmospheres post-World War II, characterized by the divergence of economic realities from the promise of the American dream. For them, affect is the felt "shared atmosphere" of crisis (p. 57): present phenomena, "under constant revision," which are "sensed" rather than apprehended "rationally" (p. 4). Here, I use affect to examine (1) a moment of crisis, transition, and uncertainty provoked by the increased visibility and success of far-right discourses as they manifest on digital platforms; (2) the "atmosphere" of job insecurity in a neoliberal, digital economy, where moderators' jobs are always on the line. Moderators are constantly encountering new ways of expressing far-right ideologies through reviewing posts. At the same time, collective fears of economic insecurity shape how they review content. Affect is a useful way to apprehend these phenomena because the concept allows us to engage with the social dimensions of moderators' uncertainties and anxieties.

Moderators negotiated these affects through stancetaking: aligning and disaligning with pieces of ideology and collective sentiments indexed by emergent vocabularies and ways of framing race on Facebook. For linguistic anthropologists, stance is concerned with how speakers and writers position themselves in relation to other social actors and discourses. As Jaffe (2009) writes, people “are necessarily engaged in positioning themselves vis-à-vis their words and texts . . . their interlocutors and audiences . . . and with respect to a context that they simultaneously respond to and construct linguistically” (p. 5). These acts of positioning have implications for how people construct their “subject positions (social roles and identities; notions of personhood), and interpersonal and social relationships (including relations of power) more broadly” (Jaffe, 2009: 4). Stancetaking is a useful concept to track the ways people negotiate sociopolitical realities, because it gives us a way to track how people navigate conflicting economic, political, and social investments. For moderators, taking up stances is a way of anchoring onto—or making sense of affective atmospheres—economic insecurity, an expanding white supremacist movement on social media—and their own values.

Crucially, moderators can become ambivalently complicit in reinforcing post-racial ideology and white supremacist conspiracy theories in the process. In linguistic anthropological terms, we might understand the act of reviewing hate speech as a “raciolinguistic” project: one that “co-naturalizes race and language in relation to longstanding histories of colonialism and nation state formation” (Rosa and Flores, 2017: 15). While Rosa and Flores (2017) are referring primarily to how Eurocentrism and white supremacy are reproduced through the distinctions we make between languages spoken by racialized peoples (hyper visible) and white people (coded as “normal”), categorizing certain content as hate speech is a raciolinguistic project in a sense that it naturalizes a post-racial frame and “mainstreams” white supremacist conspiracy theories (Miller-Idriss, 2020) in ways that make underpaid workers complicit in reproducing white supremacy.

In subsequent sections, I outline these forms of complicity. The first section of my analysis focuses on productivity metrics as mechanisms that mediate moderators’ politics, their encounters with racist content, and Facebook’s content policy. Here, I describe processes through which metrics engender complicity with post-racial ideology. The second section explores the effects of navigating an ever-shifting political landscape amid economic insecurity. For moderators, such processes produce troubling doubts about whether they, too, have become racists.

“Fucked up but not violating”

Victoria races from the parking lot into her office building to turn on a time tracker: a tool that Facebook’s third-party vendor sites use to ensure moderators are working efficiently and productively. She has a 5-minute window to log on, but the tracker has been malfunctioning all week, mistakenly reporting her and her colleagues late to work. Still, she feeds it information throughout the day: she reports her 30-minute lunch break and the time she took to get a drink of water. As she reviews potential hate speech, the tracker monitors time spent on each piece of content, warning her when she has spent too much time on a particular ticket. Later, she uses a few minutes of her wellness time to call her grandparents. This is how she centers herself: how she copes with the deluge of slurs and other hateful content that, as one of her coworkers once put it, are “fucked up but not violating.” When she returns to her desk, she reports the minutes she spent on the phone, which the tracker dutifully deducts from her 45 minutes of weekly wellness time.

When I set out to study how commercial content moderators do their work, I anticipated that moderators would focus solely on their difficulties discerning which semiotic objects violated Facebook's implementation standards (Community Guidelines as it is known to the public). Instead, moderators told me stories about feeling pressured to review complex and potentially racist content within a matter of minutes, the repercussions of a poor result on an audit, remediation classes, and having to compromise political and moral convictions for the sake of job security. These stories suggest that the material and ideological are inseparable concerns that inform how hate speech is moderated. Together with their own moral and political convictions and semiotic labor, economic anxieties are central to moderators' decision-making processes.

Translating hate speech policy into decisions such as delete, ignore, escalate—not to mention their respective subcategories—is a difficult task. Sometimes, content moderators' own cultural and historical knowledge conflicts with Facebook's policies in ways that make it difficult to remove racist content from the platform. At the same time, hateful utterances can be difficult to identify. For this reason, Victoria considers hate speech the most difficult and least straightforward type of content to identify:

> There's just so much nuance to it, and I mean, some people are blatant and upfront with what they say like, they use a slur, that's easy, but a lot of times, people are so subtle. They just hint at it. They imply things or they put up memes that you need a lot of like, historical and cultural context to understand. So yeah, there's just a lot of background knowledge that you need, especially of like, the smaller communities to kind of understand what they're saying. And all those like, dog whistles, a lot of the time that doesn‘t really fit into the policy. So even if you know what they're saying, if you can't like, prove it very literally? Based on the policy or action it and get it wrong.

The conflicting political, moral, and economic investments that content moderators must consider are well captured in Victoria's description of what it takes to moderate hate speech. While more overt speech, like slurs, are easier to moderate and are covered by Facebook's content policy, Victoria finds it difficult to adjudicate "dog whistles" and language unique to smaller communities. Her account demonstrates that moderators are at the forefront of encountering new types of hateful content amid this era of far-right resurgence.

Moreover, her labeling of coded language as "dog whistles" gestures toward two types of stancetaking: (1) her own alignment with left-leaning politics, which recognizes that contemporary racist speech in the post-racial era is dynamic and often abides by liberal democratic norms of politeness, and (2) her disalignment with Facebook content policy which aligns with post-racialism by defining hate speech in narrow terms (Ganesh, 2021; Matamoros-Fernandez, 2017; Siapera and Viejo-Otero, 2021). So, she manages "nuance"—the affective circuit of disparate, "fractious" (Stewart, 2007: 3) politics—by aligning and disaligning with certain dispositions through her use of the term "dog whistles."

The fear of "[getting] it wrong" plays an important role in moderators' decisions regarding hate speech. The need to "prove it very literally," as Victoria put it, hinges on the fear of making too many mistakes, disciplinary action, and potential job loss.

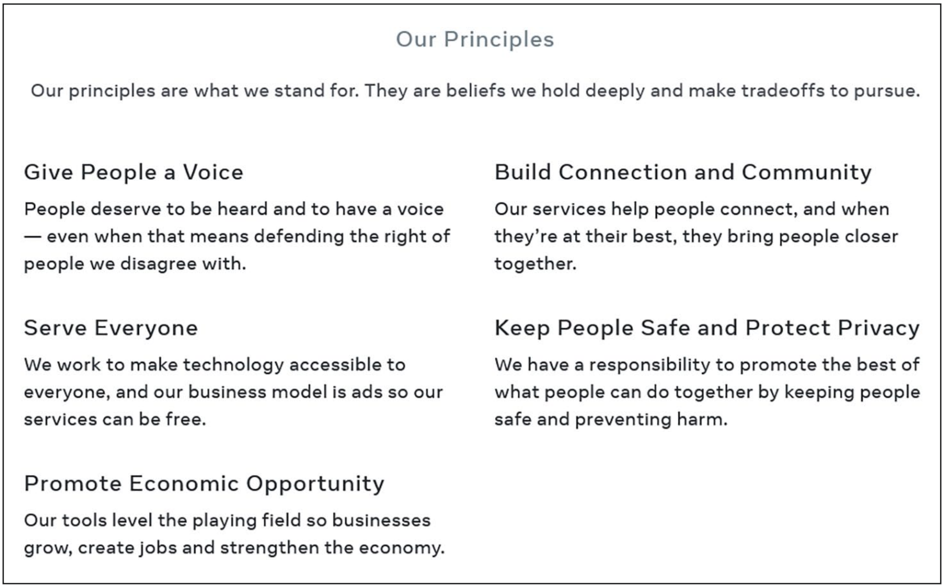
Facebook's third-party vendors keep track of moderators' mistakes through an auditing system that assigns each moderator a weekly accuracy score: the percentage of pieces of content reviewed "accurately" as determined by a human auditor. For instance, at Victoria's site, Quality Analysts (QA) calculate each moderator's accuracy score by taking a sample of 50 "actioned" (i.e. reviewed) pieces of content per week and comparing their own decisions with that of each moderator. When a moderator's decision differs from an auditor's, it counts against their accuracy score. If their accuracy scores consistently dip below 98%, they are placed in remediation programs, tested, and risk losing their jobs. As Victoria's account below demonstrates, anxieties surrounding the accuracy score and job security can conflict with moderators' own political stances and desires to rid the platform safe of racist content. It is also important to note that third-party vendors' and Facebook's rhetoric, which places the burden of platform environment on moderators, makes the task of reviewing hate speech "stressful":

> It's very frustrating. I mean—you know, you're told that like, your job is like to make the platform a welcoming place and to, um, clean it up so that everyone can enjoy it kind of thing. So, you know you're stopping harassment and um so whenever you're put in a situation like that, you're trying to think about the real-world harm but you're also thinking about your scores because that's what keeps your job, that's what like—allows you to get the schedule that you need. Um if your score had dropped. . .you would get put into like a remediation program and then from there you had to pass a bunch of tests and the tests were very like, screwed up and so a lot of people got fired from that program so it is just like, very scary if your scores drop in any way. So, it is really stressful trying to make that decision.

The statement above illustrates how moderators like Victoria weigh the possibility of job loss against their own desires to prevent "real-world harm" when reviewing hate speech. Here, she wrestles with two types of risk: the risk of harm to the community should she ignore what she identifies as hate speech and the risk of making a "mistake," that is, acting contrary to content policy, and compromising her accuracy score.

Diane also describes how she negotiated worries about getting fired with Facebook policy's political alignment and her own moral/political convictions. In this example, Diane also wrestles with her positionality as a white woman supervising coworkers who are predominantly people of color when actioning white nationalist content. Here, she troubles Facebook's "race-blind" orientation (Siapera and Viejo-Otero, 2021) toward such content but nevertheless defers to policy as written to protect herself and her team from low accuracy scores that might cost them their jobs. She told me the following story in response to Project Veritas' claim that content moderators systemically remove politically conservative speech from the platform:

> Project Veritas did that one piece about how one lady was like, undercover just deleting posts. Yeah, I don't know anyone who would have done that. Because like the reason is, is that your job is so tied to your accuracy score as an agent, and there's so much fear instilled in you about that, that it would take someone who doesn't really care about being employed . . . to start doing stuff like that. Because I remember explaining the difference between a white supremacist and white nationalist, because that's how the hate speech policy worked at the time [in 2017] is they had made the designations between one is acceptable and one is not: white supremacy is



Our Principles

Our principles are what we stand for. They are beliefs we hold deeply and make tradeoffs to pursue.

Give People a Voice

People deserve to be heard and to have a voice — even when that means defending the right of people we disagree with.

Build Connection and Community

Our services help people connect, and when they're at their best, they bring people closer together.

Serve Everyone

We work to make technology accessible to everyone, and our business model is ads so our services can be free.

Keep People Safe and Protect Privacy

We have a responsibility to promote the best of what people can do together by keeping people safe and preventing harm.

Promote Economic Opportunity

Our tools level the playing field so businesses grow, create jobs and strengthen the economy.

Figure 1. Facebook's principles.[a]
[a]Retrieved from https://about.facebook.com/company-info/. Accessed 5 October 2021.

> claiming that you're better than another race, and white nationalism is just celebrating being white, which is different, because you're not putting down. And god, I had to—I had to explain this over and over again to some people on the floor, and also, I'm a really, really white girl, obviously, and explaining that feels super wrong when like most of your floor is POC. So, there's already that dichotomy, and I'm like look, it's a shitty policy, I don't agree with it but—and I would be honest and be like, look, this seems shitty, but the reason why they're doing it is because it's a global company, and these are global policies and that if we're applying it to this, it has to be like, span things. And I get that you don't agree with it, but like we have to do it, because that's how QA [Quality Analysts] are looking at it, and that's how Facebook's looking at it.

How content moderators are monitored and assessed creates a scenario where she must choose between operationalizing a "shitty policy" or get in trouble with QA and Facebook itself. Her use of the term "POC" (people of color) to contextualize her positionality vis-à-vis BIPOC workers also indexes disalignment with Facebook policies on white nationalism and white supremacy at the time. Her use of this term indicates her recognition of the affective charge of racialized power imbalances in the workplace, which indexes a political stance left of Facebook policy's "neutral" center. Faced with a collection of disparate interests and power asymmetries, she aligns with Facebook's principles of "Voice" (see Figure 1), mobilizing its race-neutral language to manage her coworkers' concerns about leaving white nationalist content up on the platform, her own political stance, and her fears surrounding the repercussions of contradicting Facebook policy: "look, this seems shitty, but the reason why they're doing it is because it's a global company, and these are global policies and that if we're applying it to this, it has to be like, span things." By saying that the policy must "span things" in a "global company" with "global policies," she aligns with Facebook's principle of "Voice," which asserts that everyone has a right "to be heard and have a voice." In doing so, she—and perhaps her coworkers—reluctantly reproduce a "race-blind" approach to white nationalism even as they recognize that the policy is flawed.

Sara likewise gestured toward the limits of Facebook hate speech policy in capturing racist speech. Unlike content moderators like Victoria, Diane, and Diane's coworkers, who are required to action content in real time within a matter of minutes, Sara worked on training AI to identify white supremacist symbols. Still, even without the pressure to make decisions on content quickly, Sara's work was also audited, adding to the stress of her job and shaping which types of content she chose to categorize as hate speech. As she recounted,

> [Facebook was] very, very detailed [about hate speech policies], and it frustrated me sometimes, because they were so detailed that it was easy to miss a lot of things. So, you could see that something was obviously racist or um, sometimes trends would come up and the xguidelines wouldn't get around to it for a couple of months. And so, you knew something was wrong, but you couldn't do anything about it, because it didn't match the criteria of the guidelines.

I then asked her what would happen if she went against the guidelines to categorize "obviously racist" content as hate speech. She responded,

> So, you have a select number of jobs that are audited. I'm thinking maybe 200 jobs—perhaps 10 of them would get audited. So, it really kind of is luck of the draw if you get um, ones that you're audited are right or wrong. But it does kinda make it a little bit stressful to just be wondering if you're doing it correctly or not.

Here, we see again how the accuracy score constrains moderators from categorizing racist content as hate speech. For Sara, the issue is compounded by the fact that content policies were slow to respond to new "trends" in hateful content.

Rebecca points out that while Facebook content policy worked to update guidelines to match the changing repertoire of hate speech on the platform, such guidelines were slow to change. Complementing and extending Ganesh (2021), this indicates another limit of Facebook hate speech policy. Looking at how workers negotiate and deploy Facebook's policies reveals that the company's post-racial framework produces a reactive strategy toward racist content on the part of both policy workers and moderators. Without training in a framework that recognizes racism's rootedness in the structures and institutions of American society and due to pressures exerted by the accuracy score, moderators like Rebecca approach hate speech as a growing, unruly collection of semiotic objects they are powerless to manage in ways that align with their own moral instincts and political convictions.

Facebook's outsourcing structure places pressure on third-party vendors to accurately review large quantities of user-generated content. As an interlocutor once told me, content policies and moderation practices are aimed toward maintaining profitability through managing Facebook's reputation as a welcoming space that "[gives] people voice" no matter the social cost. Management in third-party vendor sites then place the responsibility of managing PR outcomes on the shoulders of content moderators through surveillance mechanisms such time trackers and routine audits that ensure moderators abide by Facebook's principles.

To understand how post-racial policy is enforced, we must examine Facebook's outsourcing model, particularly the ways in which moderators are disciplined so that they operationalize Facebook's principles, especially "Voice," no matter their personal stance. How content moderation reproduces white supremacy and enables far-right activity on Facebook is not so much a question of what moderators believe but a question of material conditions, which give workers essentially no choice but to align with—or voice (Agha, 2008)—Facebook's post-racial principles. Made to operate according to the framework that white nationalism is not problematic because it does not explicitly "put down" anybody, moderators become complicit in positioning whiteness as an identity potentially vulnerable to attack and equally deserving of protection. In doing so, they participate in a raciolinguistic project (Rosa and Flores, 2017) that reproduces what Song (2014) calls "the culture of racial equivalence": a conceptualization of racism, stemming from the growing emphasis upon racialization, "which increasingly involves multiple perpetrators, victims, and practices without enough consideration of how and why particular interactions and practices constitute racism as such" (p. 107). Moderators' experiences make clear that scholars must pay attention to the ways in which methods of surveillance, economic conditions, and other social considerations interact with post-racial ideology reflected in content policies to better understand how racist content is managed and experienced by precarious workers of mainstream, corporate social media platforms.

"I am not this person"

Facebook's structure of platform governance, which operates through the relentless surveillance of underpaid and outsourced labor, not only leads moderators to make decisions that reproduce post-racialism and enable white supremacist activity on the platform, but also adversely impacts their well-being. Because they lack training to navigate the power relations that animate US racism, they are overwhelmed by contradictory affects that surge through them all at once. Thus, understandings of racism, predicated on the idea that becoming racist is a trait one can develop, shape their sense of self and each other: they begin to doubt, questioning whether they have become racists, or their coworkers have become racists through routine exposure to racist content.

One of my interlocutors, Michael, recently got a book deal to write about his past experiences as a content moderator. I texted him one night while drafting this article and told him I was about to write about moderators fearing that they or their coworkers have become racist. He responded, "I was just writing about that yesterday, about feeling that I was becoming more racist."

While Michael declined to elaborate, he is not alone in feeling like racist content has changed him. Sara, one of the moderators from the previous section, once told me a story about hate speech symbols "[getting] into her head." As she recounted,

> Well, you know what is interesting to me um working with hate speech so much is it would get into my head. Like these images and words would flash across my brain like with no warning, and I didn't want them there. It was—it was awful. I was like, I would kind of torturing myself, um, and so I do think, you know, it can get into your head easier than you would think perhaps? Or when I was out and bout like if I ever saw a white guy with a tattoo, I would immediately start scanning his body for hate symbols. And every white guy has a tattoo in [City] . . . I was like hating myself for it. They're not supposed to be there. I'm not this person.

Elsewhere, I have analyzed Sara's story as a "haunting" brought on by the work of content moderation where the horrific and mundane converge to produce affective states. But considered within the context of reviewing white supremacist symbols, we might understand this haunting more specifically as a lingering preoccupation with "getting it right" according to Facebook's standards of hate speech. We might interpret Sara's self-hatred and her assertion that she is "not this person," as indicating her fear that working with hate speech has changed her. Or, put differently, she fears the racist alignments she witnessed have made their way into her consciousness in irrevocable ways.

Such fears are more explicit in Christine's story. In this case, she frames racism as conflict between white and Black coworkers, who have adopted "racist mannerisms" against each other from viewing certain types of content. She points to content depicting South African "white genocide" as an example of content that made her coworkers racist:

> I'm not sure if you're aware of the like the genocide that started going on in South Africa. Basically, there was a short movement that I started to see a little about before I left [Facebook] and they were talking about um retaking land from South African settlers, because initially I guess it was the Dutch who had colonized that area. And so, they were saying that they realized that more Dutch people owned land in South Africa than Africans themselves. And so, the movement that they started was literally going in and reclaiming these villages in the way that these people's ancestors supposedly did, which was going through and slaughtering and sharing the images all over the internet. So, we would see like, images of this little farmhouse or whatever and the family strung up by the rafters. There's kids, women, men it didn't matter—slaughtered and strung up.

Christine's story is a reference to a white conservative conspiracy theory that has framed South African land reform—the redistribution of land back to indigenous peoples in South Africa—as "white genocide": the far-right notion that "elites are conspiring to end the white race through immigration, miscegenation, feminism, and indoctrination" (Tebaldi, 2021: 78). For white conservative settlers, the escalation of white genocide is evidenced by the brutal murders of white farmers. But while racial vitriol does fuel the murder of some white farmers, James Pogue (2019) of Harper's Magazine reports that in 2018, "20,000 people were slain [in South Africa], most of them black." Of this 20,000, "there were only 62 farm murders." Moreover, "according to one of the country's largest agricultural associations, murders of farmers are at 20 year low. And not all of the victims are even white." Thus, Christine's story and her use of the term "white genocide" should be understood as indicative of her alignment (Du Bois, 2007; Jaffe, 2009) with a broader reactionary narrative pedaled by white conservatives who feel that their power is threatened by land reform. She continued,

> And there were groups here in America that were trying to associate with BLM that were saying, we should do that shit here and that was not getting removed, that was not getting like, because they were saying well if we try to remove it, now we're looking like we're the white supremacists, now we're looking like we're trying to block the freedom of speech or we're trying to oppress them further. And you know when they [Facebook] did start trying to cut out um that sort of racial attack on that side, they did get a lot of really negative feedback. People threatened to attack um Facebook over it because they were like, oh I see Facebook is another one of those corporate white giants trying to oppress everybody and it's like, no! If we're gonna oppress one side, from saying fucked up shit about the other, then of course we're gonna do that to the other. It's like trying to get people to understand that but that wasn't necessarily the case.

She narrates the "spread" of the migration of "white genocidal" ideology to the United States, describing how some users posted about doing the same as part of the American Black Lives Matter movement. Initially, she and her coworkers did not remove this content, fearing that if they did, they would be viewed as "white supremacists" who block oppressed peoples' "freedom of speech." But when they did start removing such content, Facebook received negative feedback from users, which she is frustrated by. For her, content moderation is about fairness and a balanced perspective of "both sides" regardless of the power imbalance between whites and Black people in the United States. Black people, she implies, are just as capable of "oppressing" whites. Again, oppression is abstracted from the history of racism and the disproportionate, systemic violence perpetrated by white people against Black, Indigenous, Latinx, and Asian American in the United States. She thus aligns with Facebook's principle of voice, described in the previous section, which neglects power differentials between white and racialized users. In doing so, she unknowingly aligns with a white supremacist narrative that justifies the oppression of BIPOC due to so-called "white genocide." She then describes how interacting with such content affected her coworkers:

> And so, we start seeing like certain people like even within the workplace, like one of the trainers, start taking on more and more like sort of racist mannerisms? And it was like you'd see some of the white people taking on a little bit more racist mannerisms, because they were feeling that aspect. They were saying like, well we're being blocked on every account, we can't say anything, like people are scared just to say, oh yeah this was a Black person because they're like, just saying that they feel like they're already like gonna be attacked, like oh you're fucking racist cause you used the word black or whatever. So, they're like, we don't even know how to move or talk about this subject anymore.

These "racist mannerisms" are based on moderators' fears that they can no longer discern between racist and non-racist behavior, practical alignments and fundamental shifts in who they are as people. If stories from the previous section show that moderators sought to manage what they understood as conflicting investments by voicing Facebook's principles, the stories in this section are moments when moderators are not quite sure where their political and moral commitments begin and end. The people they used to be, or the people their coworkers used to be, have been replaced by people who voice troubling ideologies, which pivot on different types of race-thinking. To navigate this, Christine returns to voicing Facebook's principles of voice and community (see Figure 1):

> It's like trying to keep everybody [content moderators] at a point of—remember this: we're not in this to choose sides, we're in this to try to bring people together. That's the point of this media outlet is to connect people in a healthy way that they could learn more, that they could unite and start doing something. Something positive, something good.

Through re-aligning with these principles, Christine attempts to bring her coworkers back to a space where they act as neutral observers and adjudicators of content or as conduits of disparate, global affects, who nevertheless default to Facebook policy. By training her coworkers' focus on making Facebook a "positive' space where people can "unite and start doing something," she voices Facebook's principle of community, which focuses on "[bringing] people closer together." As we have seen from the previous section, moderators' stances have material consequences. In a context where moderators' personal alignments are suppressed through mechanisms that ensure their productivity, aligning with Facebook's principles is often the only choice.

Conclusion

Through focusing on the experiences of Facebook content moderators, I have shown that political beliefs alone do not determine moderators' approach to hate speech. Instead, their decisions depend on negotiating a web of affective investments that scale up to racialized and racist ideologies. Mechanisms that discipline workers and their fears of economic insecurity often determine their alignments with certain understandings of race, racism, and Facebook's role as a platform, even if they personally disagree with such political stances. Thus, it is not so much that workers intentionally "amplify" or "censor" the voices of those with certain political stances. Rather, workers are *disciplined* in ways that leave them no choice but to align with Facebook's post-racial approach. This way of approaching hate speech can also adversely affect moderators. Acting as human conduits of affects surrounding race, racism, and the politics of speech who are made to align with a post-racial framework that erases long-standing power asymmetries, they find themselves fearful that they have become racists. In this way, moderators become complicit in a raciolinguistic project (Rosa and Flores, 2017) that upholds white supremacy through a framework that neglects hate speech's situatedness in "longstanding histories of colonialism and nation-state formation" regardless of their historical and cultural knowledge or their own political stances (p. 15).

Facebook's profit-driven, post-racial approach to hate speech is cause for concern in this political moment, given that far-right groups online and offline (Maly, 2019; Mondon and Winter, 2020) mobilize a culture of racial equivalence (Song, 2014). As Maskovsky (2017) observes, the current politics of white racial resentment hinges on a two-pronged, contradictory strategy that he calls "white nationalist postracialism": practices that "reclaim the nation for white Americans while also denying an ideological investment in white supremacy" (p. 434). Given that social media has increasingly become a space for far-right recruitment, organizing, and metapolitical activity (Stern, 2019), understanding the complex negotiations that take place when moderators mobilize Facebook's post-racial policy to "action" hate speech is crucial.

Moderators' accounts suggest that their complicity in the denial and amplification of voice depends not on targeted censorship, as the white radio show host in the beginning of this article suggested, but on Facebook's outsourcing model, which prioritizes productivity and "accuracy" over well-being and anti-racism as the moderator explained. This model creates a situation where workers align with white supremacy through post-racialism. Moreover, workers' experiences offer a window into the role of systemic racism in content moderation, which I gestured toward in the opening vignette. In this case, it is the erasure of systemic racism from Facebook's conceptualization of hate speech, coupled with oppressive working conditions, that makes moderators complicit.

Examples from moderators' experiences trouble neat distinctions between mainstream and extreme, liberal and illiberal that proliferate in both scholarly and popular publications since 2016. Thus, content moderation represents another potential "erosion zone" between liberal and far-right racism not entirely new. While Facebook moderators' realities are uniquely situated, their stories also echo the experiences of well-meaning institutional actors across Western liberal democratic societies, who do not identify as white nationalists/separatists yet are compelled to participate in the reproduction of white supremacy for reasons out of their control. For instance, Sara Ahmed's (2012) ethnography of diversity workers in British and Australian universities reveals the ways in which the language of "diversity" forecloses meaningful anti-racist restructuring in higher ed institutions. Like Facebook moderators who hope to make the platform a better place and who address racism through flawed hate speech policies, diversity workers find that institutional barriers make it impossible to realize their goals for inclusion. As Ahmed illustrates, the goal of "diversity" is to *maintain* offices and departments that specialize in "diversity, equity, and inclusion": to perform anti-racism rather than enact impactful anti-racist praxis.

Future research on racism and social media should thus situate the problem of white supremacist discourse within the long history of institutional racism by ostensibly well-meaning liberal institutions where profit is a driving force. To do so, scholars should attend to factors in the spaces of digital labor, such as the disciplinary mechanisms described in this article, which speak less overtly about people's beliefs yet facilitate troubling alignments with white supremacist projects.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was support from the National Science Foundation, grant #1928627.

ORCID iD

Rae Jereza https://orcid.org/0000-0002-3796-6154

Notes

1. This, in part, stems from difficulties associated with accessing the spaces of commercial content moderation: the task of adjudicating user-generated content on mainstream social media platforms like Facebook, Instagram, Twitter, YouTube, TikTok, Snapchat, and more (Roberts,

2016, 2018, 2019). As many have noted, this is intentional as social media companies hope to create seamless user experiences (Carmi, 2019) and avoid scrutiny by civil society and regulatory institutions alike (Gillespie, 2017, 2018). Some claim that the secrecy shrouding content moderation also arises out of concern for worker safety and fears that revealing internal processes in detail might facilitate the activities of bad actors on the platform. Such barriers prevent us from understanding what informs moderators' decision-making practices, leading some to speculate that certain groups are intentionally "censored" by moderators while others' voices are amplified. These barriers to access mean that social scientists have difficulty studying the social processes that animate moderation outcomes.

2. Recall, for instance, Mark Zuckerberg's decision in May 2020 to leave up Trump's verbatim repetition of the racist comment "when the looting starts, the shooting starts."
3. Retrieved from https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf. Accessed 8 July 2020.

References

Agha A (2008) Voice, footing, enregisterment. *Journal of Linguistic Anthropology* 15(1): 38–59.

Ahmed S (2012) *On Being Included: Racism and Diversity in Institutional Life*. Durham, NC: Duke University Press.

Berlant L (2011) *Cruel Optimism*. Durham, NC: Duke University Press.

Bonilla-Silva E (2010) *Racism Without Racists: Color-blind Racism and the Persistence of Racial Inequality in the United States*. Lanham, MD: Rowman & Littlefield Publishers Inc.

Bonilla-Silva E (2015) The structure of racism in color-blind, "post-racial" America. *American Behavioral Scientist* 59(11): 1358–1376.

Breslow J (2018) Moderating the 'worst of humanity': sexuality, witnessing, and the digital life of coloniality. *Porn Studies* 5(3): 225–240.

Carmi E (2019) The hidden listeners: regulating the line from telephone operators to content moderators. *International Journal of Communication* 13(2019): 440–458.

Chen A (2014) The laborers who keep dick pics and beheadings out of your Facebook feed. Available at: https://www.wired.com/2014/10/content-moderation/ (accessed 15 September 2022).

Cho S (2009) Post-racialism. *Iowa Law Review* 94(5): 1589–1650.

Du Bois JW (2007) The stance triangle. In: Englebreston R (ed.) *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*. Amsterdam: John Benjamins Publishing Company, pp. 139–182.

Ganesh B (2021) Platform racism: how minimizing racism privileges far right extremism,' Items: insights from the social sciences. Available at: https://items.ssrc.org/extremism-online/platform-racism-how-minimizing-racism-privileges-far-right-extremism/ (accessed 15 April 2020).

Gillespie T (2017) Regulation of and by platforms. In: Burgess J, Powell T and Marwick A (eds) *The SAGE Handbook of Social Media*. Thousand Oaks, CA: SAGE, pp. 254–278.

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

Gorwa R (2019) What is platform governance? *Information, Communication & Society* 22(6): 854–871.

Jaffe A (2009) *Stance: Sociolinguistic Perspectives*. Oxford; New York: Oxford University Press.

Jereza R (2021) Corporeal moderation: digital labour as affective good. *Social Anthropology* 29(4): 928–943.

Maly I (2019) New right metapolitics and the algorithmic activism of Schild and Vrienden. *Social Media + Society* 5: 1–15.

Maskovsky J (2017) Toward the anthropology of white nationalist postracialism: comments inspired by Hall, Goldstein, and Ingram's The hands of Donald Trump. *HAU: Journal of Ethnographic Theory* 7(1): 433–440.

Matamoros-Fernandez A (2017) Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook, and YouTube. *Information, Communication, and Society* 20(6): 930–946.

Messenger H and Simmons K (2021) *Facebook Content Moderators Say They Receive Little Support, Despite Company Promises*. Available at: https://www.nbcnews.com/business/business-news/facebook-content-moderators-say-they-receive-little-support-despite-company-n1266891 (accessed 18 April 2022).

Miller-Idriss C (2020) *Hate in the Homeland: The New Global Far Right*. Oxford; Princeton, NJ: Princeton University Press.

Mondon A and Winter A (2020) *Reactionary Democracy: How Racism and the Populist Far Right Became Mainstream*. London; New York: Verso Books.

Newton C (2019) *The Trauma Floor*. Available at: https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona (accessed 18 April 2022).

Pogue J (2019) *The Myth of White Genocide*. Available at: https://pulitzercenter.org/stories/myth-white-genocide (accessed 13 February 2021).

Roberts ST (2016) Commercial content moderation: digital laborers' dirty work. In: Noble SU and Tynes BM (eds) *The Intersectional Internet: Race, Sex, Class and Culture Online*. Bern: Peter Lang Publishing Group, pp. 147–159.

Roberts ST (2018) Digital detritus: "Error" and the logic of opacity in social media content moderation. *First Monday* 23(3). Available at: https://firstmonday.org/ojs/index.php/fm/article/view/8283

Roberts ST (2019) *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, CT; London: Yale University Press.

Rosa J and Flores N (2017) Unsettling race and language: toward a raciolinguistic perspective. *Language in Society* 46(5): 621–647.

Siapera E and Viejo-Otero P (2021) Governing hate: Facebook and digital racism. *Television and New Media* 22(2): 112–130.

Song M (2014) Challenging a culture of racial equivalence. *The British Journal of Sociology* 65(1): 107–129.

Stern AM (2019) *Proud Boys and the White Ethnostate: How the Alt-right Is Warping the American Imagination*. New York: Penguin Random House.

Stewart K (2007) *Ordinary Affects*. Durham, NC; London: Duke University Press.

Tebaldi C (2021) Make Women Great Again: women, misogyny, and anti-capitalism on the right. *Fast Capitalism* 18(1): 71–81.

Author biography

Rae Jereza is a queer Filipinx anthropologist living in occupied Nacotchtank land. At the time of this article's acceptance, they were a postdoctoral researcher at the New Jersey Institute of Technology. Currently, they are a Senior Researcher at the Polarization, Extremism and Research Innovation Lab and Research Assistant Professor at the School of Public Affairs at American University.