Trans-Centered Moderation: Trans Technology Creators and Centering Transness in Platform and Community Governance

Hibby Thach
hibby@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Samuel Mayworm
mayworms@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Michaelanne Thomas
mmtd@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Oliver L. Haimson
haimson@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

ABSTRACT

Mainstream platforms' content moderation systems typically employ generalized "one-size-fits-all" approaches, intended to serve both general and marginalized users. Thus, transgender people must often create their own technologies and moderation systems to meet their specific needs. In our interview study of transgender technology creators (n=115), we found that creators face issues of transphobic abuse and disproportionate content moderation. Trans tech creators address these issues by carefully moderating and vetting their userbases, centering trans contexts in content moderation systems, and employing collective governance and community models. Based on these findings, we argue that trans tech creators' approaches to moderation offer important insights into how to better design for trans users, and ultimately, marginalized users in the larger platform ecology. We introduce the concept of *trans-centered moderation* – content moderation that reviews and successfully vets transphobic users, appoints trans moderators to effectively moderate trans contexts, considers the limitations and constraints of technology for addressing social challenges, and employs collective governance and community models. Trans-centered moderation can help to improve platform design for trans users while reducing the harm faced by trans people and marginalized users more broadly.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**.

KEYWORDS

content moderation, transgender, platforms, online communities, governance, trans technologies

ACM Reference Format:
Hibby Thach, Samuel Mayworm, Michaelanne Thomas, and Oliver L. Haimson. 2024. Trans-Centered Moderation: Trans Technology Creators and Centering Transness in Platform and Community Governance. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3630106.3658909

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658909

1 INTRODUCTION

Trans[1] people frequently experience challenges when interacting with online platforms and their moderation processes [16, 26, 36]. Often, platforms' moderation systems and policies result in excluding trans people from the platform. Examples include platforms guidelines related to gender that do not acknowledge trans users [4] and algorithmic moderation tools that inaccurately assess and disproportionately target trans users' content [4, 26, 45]. Excluding trans people from online platforms via moderation processes exposes them to harms, such as reduced access to community, support, and public speech. This exclusion hinders trans people's ability to use online platforms as freely as cisgender people or to address their trans-specific needs, such as community building, information seeking, and mutual aid.

Trans tech creators[2] design and build technologies specifically for trans individuals, including online spaces and communities. These trans technologies[3] are tailored to keep trans users safe while enabling them to meet their unique needs [27, 29]. Trans technology design contrasts sharply with mainstream technologies and platforms, which often overlook trans people in their design considerations [4, 27, 29, 38, 40]. Unlike mainstream platforms, trans technologies frequently adopt content moderation systems and approaches that prioritize the safety of trans users, placing explicit focus on addressing their specific concerns. Additionally, talking to people who *create* trans technologies (rather than technology users) uniquely makes visible the substantial work that moderation requires.

We examine trans tech creators' design processes and considerations for designing content moderation systems that prioritize

[1]We use "trans" in this work to refer to a wide range of transgender experiences, explicitly including nonbinary trans people.
[2]By "trans tech creators," we mean people who created trans technologies; these creators were not necessarily trans themselves, though about 80% of them in our study were.
[3]Haimson et al. describe trans technology as "technology designed specifically to address some of the challenges trans people face in the world, often designed in response to the lack of representation in more mainstream technologies" [27, 29], using a practical definition of "trans technology" rather than a more theoretical definition [25].

trans needs, experiences, and identities. We asked: (RQ1) What challenges do trans tech creators encounter with content moderation? (RQ2) How might trans technologies offer possible solutions to these moderation challenges?

To answer these questions, we conducted 104 interviews with 115 trans tech creators about their technologies and design processes. We asked creators to describe how they designed moderation processes and systems that prioritize trans users' needs, experiences, and identities, along with challenges they encountered while designing and employing those processes. We found that trans tech creators designed their moderation processes with trans users' experiences on mainstream social media platforms in mind. They addressed these challenges by carefully vetting users, appointing trans moderators, and implementing collective governance and community-driven governance models, aiming to moderate in ways that center trans users' needs and safety.

We discuss how trans-centered content moderation solutions are more likely to be employed in smaller-scale trans-built communities; yet this puts the onus on trans users and trans tech creators to moderate themselves, since trans-centered content moderation principles cannot be easily deployed on larger mainstream platforms. We argue that trans tech creators' content moderation design decisions offer important insights on designing better moderation systems that are inclusive of trans people. Further, we argue that mainstream platforms should draw from trans technologies' smaller-scale trans-centered content moderation implementations to improve their own content moderation systems. Doing so could allow mainstream platforms to more fairly moderate trans users' content and better protect trans users, and ultimately, marginalized users more broadly. Additionally, smaller-scale trans-built communities are sometimes interconnected across many spaces and platforms, offering a possible solution to the scale problem that context-centered content moderation approaches often encounter [9, 22].

Past work has highlighted trans social media users' experiences with online content moderation, particularly their exclusion from mainstream platforms' content moderation systems [26, 36]. We extend this work by contributing an empirical understanding of how trans tech creators design content moderation systems for their own technologies, highlighting how their content moderation design decisions center trans users' safety on their communities and platforms. We contribute the concept of *trans-centered moderation*, defined as content moderation systems and approaches designed to:

- review and effectively vet transphobic users and content
- appoint trans moderators to effectively moderate trans contexts
- consider the limitations and constraints of technology for addressing social challenges
- employ collective governance and community models

Trans-centered moderation as an approach deals with the intricacies of trans contexts, but offers insights into how to better consider other marginalized contexts as well – by listening to marginalized technology creators and users to understand what types of systems, policies, and platform governance work best for their communities.

2 RELATED WORK

2.1 Trans Technologies

Though trans technologies vary greatly in design and purpose, they share similar goals related to helping trans people meet their trans-related needs [27]. Some are designed to help trans people find support and develop community, such as trans-centered Discord servers like Trans Peer Network (an online peer support community for trans people) or discussion forums like Susan's Place. Many trans technologies are designed to either provide trans-related information resources for trans people (such as the Trans Language Primer, an educational resource for trans-specific vocabulary) or to connect trans people to real-world trans-friendly resources (such as Erin Reed's Informed Consent Map, a map of trans-friendly medical providers who provide informed consent HRT in the United States). Another common trans technology category connects trans users to trans-related safety information and resources, such as Erin Reed's Anti-Trans Legislative Risk Map (a map indicating which US states face the highest risk of passing anti-trans legislation) or U-Signal [52], a prototype designed to help trans people of color quickly message contacts if they experience danger in public. Many trans technologies are designed to facilitate trans self-expression based on identities and life experiences; these include art and games such as *Validate*, a dating sim game that centers queer, trans, and nonbinary people of color's narratives. Technologies can also sometimes be "trans" even if they were not designed specifically for trans people – for example, Tumblr, a social media platform that was once heavily used by trans users for self-expression, information-seeking, and resource-sharing [25]. However, Tumblr can no longer be considered a trans technology due to its 2018 policy changes banning "female-presenting nipples" and other forms of nudity and sexual content, functionally censoring many trans art and health-related resources [25, 55]. We build on past literature by examining how some trans tech creators design content moderation systems, often by centering trans safety needs and employing collective governance models that may differ from major platforms' disproportionate removal of trans users' content.

2.2 Content Moderation and Trans Users

Trans people rely on social media to meet some of their trans-specific needs, such as seeking trans healthcare information [1], crowdfunding for gender-affirming healthcare [2, 19], expressing trans identity [16], and finding community [7, 24, 47, 50]. However, past work has found that trans social media users disproportionately experience the incorrect removal of their social media accounts or content, even when they have not violated platforms' policies [12, 16, 17, 23, 26, 28, 45, 51]. Trans social media users are particularly likely to have content featuring their bodies algorithmically incorrectly flagged as "explicit" and removed [26]. Trans social media users also experience incorrect reporting of their content by other social media users as a form of transphobic harassment, exacerbating disproportionate content removals [31, 44].

Despite the importance of content moderation systems as tools that remove harmful or illegal content, many platforms' content moderation systems are designed in a way that enables discriminatory moderation practices against trans users [26]. Major platforms like YouTube [43], Instagram [5], and TikTok [12, 16] have faced

criticism for algorithmically suppressing content posted by trans users. Some trans users fight back against this algorithmic suppression; as one example, transfeminine TikTok users often adjust their posting behavior to prevent their trans-related videos from being suppressed by the platform [16]. Disproportionate removals of trans users' content harms trans users while limiting their ability to use social media for self-expression, activism, information-seeking, or other trans-related needs, ultimately preventing trans people from using social media as freely or safely as cisgender social media users [23, 26, 45]. Our research expands on past work by centering the trans user in content moderation approaches, considering how best to govern our platforms based on what may help reduce harm for trans users.

2.3 Community-Based/User-Driven Moderation

Seering [49] describes "community-based" moderation as a form of user-driven moderation where the majority of content moderation tasks are performed not by employees of the platform, but by platform users themselves. Though most major social media platforms employ a centralized, "top-down" commercial content moderation model [49], platforms like Twitch, Reddit, and Wikipedia instead employ "bottom-up" user-driven moderation [6, 49, 54]. Community-based moderation tasks vary by platform but involve regulating user behaviors, enforcing community guidelines, and removing harmful content [8, 32, 41, 49, 54]. Community-based moderation is typically driven by groups of volunteer moderators who often do not have formal training in content moderation practices, instead drawing their moderation philosophies from past moderation practices in their communities, their own personal experiences, and conversations with other moderators and users [49, 54].

Past literature has explored user-driven content moderation dynamics in online spaces that center marginalized users and communities. Thach et al. [54] explored the "unique user-moderator dynamics" on /r/FTM (a large Reddit support community for transgender men and transmasculine individuals), describing how its moderators seek community input on guideline changes by directly interacting with community members, a strategy that also helps them combat Reddit's "toxic technoculture" [34] stay safe during waves of sitewide transphobic abuse. Wu and Semaan [57] explored how Reddit's algorithmic moderation tools often fail to detect "color-blind" and covert racist content, which can obstruct volunteer moderators' efforts to moderate racist content. Gilbert [21] described how the moderators on /r/AskHistorians (a large history-themed Reddit community) employ an alternative model for proactive, justice-based content moderation, while introducing "intersectional moderation," a model for content moderation drawing from Black feminist theory that "accounts for the impact of power across multiple levels of domination and areas of resistance to oppression." Seering [49] suggested that identity-based social media user communities in more centralized moderation environments (such as Black Twitter users organized around #BlackTwitter) could potentially benefit from community-based, user-driven moderation models. However, Seering [49] also noted that community-moderated spaces are not inherently safe for marginalized communities, because user-driven moderation models are also sometimes employed by online extremist and hate-group communities. This paper builds on past work by analyzing what trans communities are currently doing online and how their collective governance and community models can inform content moderation design.

3 METHODS

3.1 Data Collection

We conducted 104 interviews with 115 trans tech designers and creators in 2021 and 2022. Using criterion sampling [35]), or selecting participants who met particular predetermined criteria, we gathered data from creators, designers, and developers of many different types of trans technologies (e.g., apps, social media sites, websites, online communities, etc.). Additional inclusion criteria included speaking and understanding English, and being at least 18 years old. We recruited participants by creating a list of potential trans technologies that drew from several years of observing the trans tech landscape, as well as by systematically searching app stores and search engines for key terms, including "transgender," "transgender technology," and "transgender apps." We continued to expand the list through snowball sampling by asking interviewees to recommend other trans technologies or trans tech creators. Categories of technology in our dataset were as follows: streaming, crowdfunding site, hackathon, online community, podcast, supplies, art, appearance-changing technology, browser plugin, body technology (e.g., prosthetics, biohacking), safety technology, mixed reality, resource site, transition app, archive or database, game, social media, dating app, voice technology, and health resource. We contacted participants via email or social media to invite them to participate; our response rate of completed interviews was 43.7%. We conducted semi-structured interviews via Zoom lasting approximately sixty minutes (mean = 63 minutes, standard deviation = 14 minutes, range = 33-93 minutes). We asked participants about the ideation and creation of their technologies, the design processes and who was involved, challenges they faced, their conceptions of trans technology, and more. With a semi-structured format, interviews focused on topics most salient to participants. Participants were compensated with a $100 gift card or check. This study was reviewed and deemed exempt by our university's Institutional Review Board.

3.2 Data Analysis

We recorded interview audio and later transcribed them for data analysis; data analysis was conducted alongside data collection. We iteratively adapted our interview protocol based on what we learned through analysis. We began by open coding [11], drawing out major themes such as ways trans people designed content moderation systems for their technologies. Through an iterative coding process, we developed themes that we continued to revisit and refine. In this paper, we focus on codes related to trans technologies and content moderation; as such, we only discuss trans technologies that include some form of content moderation system or tool (whether existing or potential). Following open coding, we then engaged in axial coding to group the codes into larger categories [11]. Our most prevalent themes, transphobic abuse, disproportionate content moderation, careful moderation and vetting, centering

trans contexts in content moderation, and collective governance and community models, are our primary focus areas in this paper.

4 RESULTS

In this results section, we describe trans technology design processes alongside considerations for content moderation processes that prioritize trans safety. Describing themes in interview participants' responses[4], we detail how trans tech designers described their thoughts about and experiences with content moderation processes across social media in relation to trans safety. Specifically, we detail 1) challenges trans tech creators encounter with content moderation and 2) how trans technology offers possible solutions to these challenges.

4.1 Challenges trans people encounter with content moderation

Like trans users more broadly, trans tech creators face challenges when interacting with different platforms and technologies' content moderation processes as related to their trans technologies. Specifically, trans tech creators described themselves and their users experiencing transphobic abuse and facing barriers to expressing their identities online because of rigid technological systems. Detailing these challenges, we show the unique difficulties trans tech creators face from content moderation processes when navigating the Internet and governing their own spaces.

4.1.1 Transphobic abuse. Trans social media users experience high rates of transphobic abuse and harassment on digital platforms [26, 47]. Gwendolyn Ann Smith, a trans technology pioneer, recalled experiencing *"substantial amounts of abuse on big platforms [such as] Twitter and Facebook,"* stating that trans social media users "don't see great moderation on those services;" Smith also shared her perception that *"Twitter does not always want to deal with users abusing trans people on their service."* Laura Horak, project leader for the Transgender Media Portal, a collaborative online database of trans filmmakers and their works, described potential forms of transphobic trolling and abuse that her team considered during the Portal's design process:

> *There's so many kinds of trolling... people are very creative when it comes to trolling. So we want to benefit trans people by trying to anticipate that, and by not exposing [trans users] to online harm and abuse. And trying to navigate that has been really hard, or just challenging to think about.*

Not only is navigating transphobic abuse complicated, but it also requires constant vigilance. *"The time to transphobia is measured in seconds, not days,"* said Jaylin Bowers, part of the team behind Trans Family Network, a network dedicated to connecting allies across the country with trans people and their families in need of support. *"It's not even people who are hostile to trans people, it's anyone who wants to take advantage of a large community. Essentially, we're a big database full of people who are vulnerable and need things."* Though large online communities are likely to be targeted by trolls, whether transphobic or not, Trans Family Network users are especially vulnerable to transphobic trolling and abuse. An organization like Trans Family Network is powerful in that it brings together many trans people and allies online to facilitate support exchange, but at the same time, it is vulnerable because of its size.

Transphobic abuse can also manifest in more insidious ways, such as when a harmful person infiltrated Trans Peer Network, a Discord trans peer support community. While many trans technologies maintain their own technological platforms, some like Trans Peer Network rely on larger platforms like (in this case) Discord. Trans Peer Network co-creator Laur Bereznai recalled, *"Two years ago, for Trans Day of Visibility, we had somebody who contacted us and claimed to be an activist specifically in a certain area and was very active, and we were happy to have them on board."* Later, however, *"we were told that this person is going by a different identity and is a serial abuser, and somebody who has been defrauding trans organizations for a long time, and has been doing a lot of really nasty shit."* Although this person was eventually removed from the Discord server, Bereznai faced substantial blowback from their community for failing to vet this bad actor well enough. Community members lost trust in Trans Peer Network's moderation team and leadership, and moderators had to deal with this blowback, despite this being a larger issue regarding online transphobia. As this example demonstrates, not only does transphobic abuse happen in obvious ways, it can happen in more concealed ways by intentionally bad actors that are hard to navigate for both community members and moderators.

Rosa Chapperri, a member of the Transverse, a media network, online community, and resource hub for transgender, non-binary, intersex, and gender non-conforming people, also spoke to us about how harassment continued even after banning troll accounts or removing transphobic content. She said, *"They [trolls] do like to just create new accounts all the time. So, it's just a continual loop of them creating new accounts and us banning them."* With all of this in mind, we see how transphobic abuse on different platforms and technologies is a complex and constant problem for trans tech creators and trans online community moderators who are trying to maintain safe environments for their communities. Having to confront, mitigate, and navigate around transphobic abuse to protect trans users is difficult, and larger platforms often offer little substantial help to alleviate these issues.

4.1.2 Disproportionate content moderation. In addition to transphobic online abuse, trans tech creators also must grapple with rigid technological systems that hinder their and their users' abilities to express themselves, or algorithms and policies that target them disproportionately [26, 54]. *"I was in the process of coming out, and... I ended up with one of the AOL 'sign on for five hours free,' pre-cassette diskette... I found some other trans people that were on the system and started a relationship."* said Gwendolyn Ann Smith (creator of Transgender Community Forum and the Remembering Our Dead website). *"You could find like-minded people... I found that fairly enticing, not only for myself, but if looking at this would interest me, then there's gotta be other people out there... who are trying to find themselves and who could use this."* She continued to tell us about how AOL at that time did not allow discussion of transness on their service, so she first had to challenge those policies before starting the Transgender Community Forum AOL

[4]We refer to most participants and their technologies by names throughout our Results section because most stated that they wanted to be identified.

online community. *"[In the 90s] on AOL, you could get kicked off the service for violations of their terms of service for using the word transsexual.... So, we used to have to do a lot of things,"* says Andrea James, creator of Transgender Map, who was also part of AOL trans online communities in the 1990s. Because they were not able to use the word "transsexual," *"we would call chat rooms 'The Gazebo,' because everybody would be able to find that, but it was an unusual enough word that other people wouldn't be using it."* Though newer platforms have replaced AOL for trans online communities, this is just one example of how trans users navigate around and resist policies that restrict their access to community and self-expression, and how community organizing around trans online censorship dates back decades. Smith was eventually successful in convincing AOL to change their policy and to allow people to discuss trans issues on the platform [13, 33], but sadly, even with community advocacy, policy change does not always occur.

Trans tech creators continue to face disproportionate content moderation on contemporary social media platforms. For instance, Guerrilla Davis of Arm the Girls, a mutual aid effort in the Bay Area that equips Black and brown trans femmes with self-defense tools, told us about the group's challenging experience with shadowbanning on Instagram:

> *Our account is shadowbanned on Instagram, and we've been repeatedly flagged for "inciting violence." Because a lot of our marketing campaign has weapons, and has guns, even though we're not necessarily promoting guns or promoting gun usage. We just wanted to start a conversation about what violence and what safety looks like.... Because if you go to the Army, or the Coast Guard, or any military Instagram page, everyone has guns, they're bombing people, there's people in uniform with assault rifles and all these things. But that's still normalized, and I'm pretty sure that those accounts aren't shadowbanned.*

Davis suspected that Instagram's algorithm shadowbanned content of trans femmes with guns – content that is explicitly trans, but also contains "adult" or "violent" content. Yet military Instagram pages do not seem to face the same scrutiny. In the case of Arm the Girls, Black and brown trans femmes are arming themselves for self-defense from the violence many trans femmes of color face. By disproportionately removing "violent" content from trans accounts but not military accounts, Instagram's algorithm may be disproportionately censoring trans bodies, language, and politics. This is not a unique case, however, as other participants expressed similar sentiments about censorship and shadowbanning. *"It's [social media] run by bots; it's been proven that they are homophobic and transphobic... we've had posts that are pulled down,"* Scout Rose from Transguy Supply, a trans-owned online marketplace dedicated to supporting trans men, trans masculine and non-binary people, said to us. *"You can contest it, but... they're not actually going to put a pair of eyes on it.... I mean, we get our posts pulled every now and then."* For context, part of Transguy Supply's goods are "packers," penis-shaped products that help to fill in crotch space in clothing for those who use them, who are often trans men or other gender-diverse individuals. As Rose told us, *"if it looks like a penis, to Facebook it's a penis, and you're selling adult products."* Despite Facebook's guidelines allowing photographs, paintings, sculptures, and other art or figures of nudity, Transguy Supply's content is still taken down occasionally, despite following community guidelines.

When these mistakes and disproportionate moments of content moderation happen, they can be difficult and time consuming to appeal. *"I've been very wary of social media because I've been banned multiple times, especially YouTube,"* said Alex of Transthetics, a company making prosthetics for trans men and other gender-diverse individuals, *"and then it's taken me three months to get it [my channel] reinstated because you just can't talk to an actual human in places like that... you're just dealing with bots the whole time."* Both Rose and Alex sell packers or various trans prosthetics that often get flagged as "adult," despite following site guidelines. They also both expressed how infrequently content moderation appeals succeed for trans tech creators, as platforms rarely allow users to interact with human moderators. Instagram has a history of denying shadowbanning and responding by elaborating on how their algorithm works, acknowledging that posts categorized as "inappropriate" would not be featured on Instagram's Explore page, despite being within community guidelines [20]. Many users still report that shadowbanning happens, and this remains a gray area that requires users to create algorithmic folk theories to describe their experiences [15, 46]. Disproportionate content moderation burdens trans tech creators with additional labor and reduced engagement and sales, and creates added barriers for trans tech creators and the people who use their technologies.

4.2 Solutions trans tech offers for content moderation

In response to the challenges trans tech creators and their technologies' users encounter with content moderation processes, interviewees discussed their existing and proposed solutions for helping to alleviate these issues. Potential solutions include 1) carefully moderating and vetting users and content, 2) centering trans contexts, and 3) collective governance and community models.

4.2.1 Careful moderation and vetting. The dominant culture of abuse and lack of care for trans users influenced how some trans tech creators designed content moderation systems for their own technologies. Taylor Chiang described the need to design a system that allows moderators on TranZap (an app for trans people to describe their experiences with physicians or to read other users' reviews) to preemptively manually review submitted comments, with the goal of avoiding publishing transphobic comments on the platform:

> *We're implementing a system in our backend database where users will submit a review, but they are not immediately published. Instead, [the comments] get put in this holding database, and then we manually look at the review, make sure that there are no transphobic comments, or just general spam – things that are not useful to the community.*

Similarly, developer Justin Bantuelle said they were *"very mindful of the fact that there are some pretty horrible people out there and really restricting the ability for anyone to do anything hateful [with the website] was a big focus."* The technology they developed, the

Gender Infinity Resource Locator, supplies online visitors with a network of providers that offer healthcare, consultation, clinical services, and additional resources focused on gender affirmative care. *"I gave them [Gender Infinity Resource Locator moderators] a full administration tool, so that if somebody leaves, if somebody becomes a bad actor, for whatever reason, they can remove their access,"* Bantuelle stated. In both Chiang and Bantuelle's examples, we see how trans tech designers view transphobic abuse as being frequent enough to figure prominently in their content moderation processes.

Beyond carefully vetting content, several trans tech creators also discussed how they designed their technology to mitigate the risk of abusive users joining their services with the intention of harming trans users. Jaylin Bowers spoke to us about the risk of transphobic people joining Trans Family Network to abuse or exploit particularly vulnerable trans users. *"We don't want to allow [abuse] to happen on the platform,"* said Bowers. *"We don't want to connect people in need with people who are predatory."* She described the process of manually reviewing potential users and assigning them a *"trustworthiness score"* before deciding whether to allow them to join Trans Family Network. *"We generally vet [potential users] on social media. The reason we include long-form answers to questions and indirect questions about our value statement and [users'] past experiences... is [because] you can suss out most disingenuous people through those answers."* Bowers described the process further: *"we go through their onboarding questions, their long-form answers, their social media... then we assign [the user] a score based on that. We have some guidelines and guidance for the trustworthiness score that we're talking about."* In this example, rather than using subjective vetting criteria, Trans Family Network uses a quantitative metric based on how much information incoming users verify to decide if a potential user is safe to let into a space, as a way to protect trans users from harm. Calculating a quantitative "trustworthiness score" may miss some of the complexities necessary for vetting, and shares some elements with algorithmic content moderation as employed by mainstream platforms – which we know are not always particularly inclusive. However, quantitative vetting is an interesting approach worthy of further research to see how it may best support trans-centered moderation.

Rosa Chapperri (from The Transverse) described Aegis, an innovative collaborative vetting system she developed that provides security for many different LGBTQIA+ safe spaces on Discord and content creators on Twitch. Moderators from each community can use Aegis to report anti-trans and anti-queer trolls and other harmful users. Aegis was developed in response to the collapse of a similar system called Pride Shield, *"where people shared usernames of trolls and... servers that were involved in Pride Shield would also ban those people from their own servers.... One community would send out a warning and everyone else would get it."* Chapperri created Aegis to recreate and improve upon Pride Shield. Elaborating on Aegis's vetting process, Chapperri said:

> *So, trolls and their usernames could be shared into a specific room [on Discord], then I programmed bots, which grabs all the information from those accounts that are being reported. I can make it show when their account was created, their username, user picture so people can recognize them, and their ID codes and everything. People can post the reason why they're banning the certain person. That gets announced to all other Discords that are following. They can ban them either on their Discord or their Twitch accounts, which in the future will probably expand to YouTube accounts, and probably some other things...*

As these examples illustrate, trans tech creators use proactive (Trans Family Network) and reactive (Aegis) vetting mechanisms to protect trans users from abuse. Both cases suggest that having human moderators manually review content may offer advantages for content moderation on trans technologies, as automated systems may overlook or misclassify disingenuous responses to a screening survey and harmful actors in an online community [9]. The examples in this section highlight ways to prioritize trans safety when designing content moderation processes.

4.2.2 Centering trans contexts in content moderation. Considering trans safety in designing content moderation processes requires centering transness instead of employing a context-agnostic "one-size-fits-all" approach that many larger social media platforms use [9, 48]. To center marginalized groups when moderating content, Haimson et al. [26] argued that directly involving members of marginalized groups while developing content moderation policy helps consider that group's context while "reducing content moderation disparities." Some participants in our study spoke on this and discussed ideas around hiring marginalized groups as moderators.

One participant who wished to remain anonymous, a Product Designer for a queer social and community app, spoke to us about trans vs. cisgender moderators for trans users and content. *"We have a community moderator who is also trans and so they're able to look at specific cases as someone who's been blocked or reported and come at it from a trans angle too,"* they said, *"which I think is super useful to not have a cis person looking at these very delicate situations and what is or isn't appropriate in the trans community."* However, even if a trans moderator may have more knowledge and exposure to trans politics and language, there can still be intra-group conflict. For example, Laura Horak (of the Transgender Media Portal), said, *"there's going to be disagreements between the trans community too, where there's not necessarily one right answer around language."* Like any other identity group, trans people have similarities to one another, but are not monolithic and have varied interests, beliefs, politics, and additional salient identities. Yet because many trans people face a common set of challenges when using social technologies, trans moderators can understand trans contexts more easily than cisgender moderators can. Thus, a trans-centered content moderation approach would acknowledge the benefits and strengths of enabling trans people to moderate trans content.

Whether a technology employs trans moderators or not, its designers must consider how some technologies can be pervasively anti-trans if they exclude or do not consider trans contexts. For example, some platforms require legal names and do not adequately provide users agency in choosing a preferred name to be displayed instead of a deadname (a name no longer in use). Willow Hayward created Deadname Remover, a Google Chrome extension that removes and replaces deadnames with chosen names, in response to witnessing a trans loved one's experience using a web browser.

"He'd click the, 'Okay. Submit assessment.' It would be like, 'Hi, deadname.' He wasn't prepared for it... and once it's happened a few times unexpectedly, suddenly you're always expecting it, even if it's not going to come up." Tools like Deadname Remover help trans people avoid the stress and discomfort that can arise when signs of their discarded identity pop up unexpectedly in their day to day online lives. Many other similar plug-ins exist, such as Dev (Github username: Derwentx)'s Jailbreak the Binary or Sophie Debs' Gender Neutralize[5]. Trans-centered moderation involves solutions like these that consider how best to reduce technological harms for trans people.

Trans technologies step in when mainstream technologies fail to fully consider trans contexts. While web browsers can resurface users' deadnames, trans-centric browser extensions cover them back up. Trans-centered moderation approaches understand, prioritize, and adjust trans contexts to mitigate the harms trans people often face online.

4.2.3 Collective governance and community models. Several trans tech designers expressed a desire to implement more collective or democratic governance structures in their technologies' content moderation systems instead of emulating the top-down [49] forms of governance found on mainstream social media platforms. By collective governance, we refer to governance systems that involve community members in the governance process as equals to creators, different in that choices are made considering the collective's views, rather than a handful of volunteer moderators. Manali Desai of Flux, a prototype social media and transition app, expressed that their desire to develop a democratic governance structure for Flux came from *"wanting to prevent harm"* toward their trans users. Specifically, by avoiding *"falling into the pitfalls of other social media [platforms]"* that employ more top-down governance models. Top-down governance models often employ content moderation tools that are efficient and can moderate a large user-base, but ignore context and often end up further marginalizing their marginalized users [48, 54]. For example, as Delilah D'Lune (creator of various trans technologies including games, social media, and an online porn platform) told us, platforms make *"governance choices that are often not in the interests of sex workers, leading to a great deal of precarity."* Because many trans tech creators had experienced mainstream platforms disproportionately moderating trans people and other marginalized groups, it is understandable why many participants in our study advocated for collective governance on smaller bottom-up [6] platforms. Yet trans technologies' governance models and their designers' moderation strategies vary, reflecting the differing needs and goals of different types of technologies and their users.

"I've been wanting to... build a platform that is collectively governed by its members for a long time," said D'Lune. *"What I wanted to do differently here is, have a truly open collective where literally everyone who is participating in the organization is automatically able to participate in the governance to create proposals, discussing them and voting on them."* She discussed the natural tendency for a majority group to appear in such a collective and how she would implement a representative policy, enabling people from smaller demographics to hold equal say as majority groups. Laur Bereznai of Trans Peer Network, spoke about a similar process, saying, *"What is really important for us is... the idea of continuously regenerating and reshaping the community as a space that it supports and empowers its members by limiting hierarchy, hierarchical structures, and centering the margins."* They talked about the team behind Trans Peer Network being more like *"plumbers than CEOS,"* making *"sure things work"* and that *"if something breaks,"* they can fix it; their community is a *"garden"* that they all tend together. In these examples, collaboration and community are prioritized rather than hierarchies, and power is distributed or renegotiated to intentionally center and uplift the most marginalized.

Community and collective accountability are central to governance conversations. *"I think that you need community. I think it's very easy to end up with situations where you have one person who's in control of a resource that should be a common resource,"* says Hayward (of Deadname Remover), *"Community is really important to make sure that no individual can take control of something, that it remains in the hands of those who need it, of those who will use it."* Similarly, Riley Johnson from now-defunct health information resource site RAD Remedy, discussed the site's governance: *"we were collaboratively, collectively operated... in some organizations that have come later, it's this person's way and that's it. We were purposeful about not doing that."* Many trans tech creators valued the contributions of their collaborators, users, and communities, and sought their input in meaningful and substantial ways. Even when trans technologies were not designed in fully collaborative ways, their deployment often stretched across multiple communities, like with Aegis's vetting system. Collective governance and community-driven models are central to many trans technologies, and may even offer possible solutions to the problem of scale (e.g., Aegis's system spanning many small communities) when it comes to transphobic abuse and counteracting issues from disproportionate content moderation.

5 DISCUSSION

We described the content moderation challenges trans tech creators face, and some possible solutions trans technology offers to these challenges. Transphobic abuse and disproportionate content moderation may be partially addressed by centering trans contexts in content moderation, and by implementing collective governance. Both approaches help center transness and marginalization when it comes to content moderation decisions; however, these fixes are more likely to happen on a smaller scale (e.g., on trans-built platforms and communities), which does not account for the larger platform ecology. Platforms' reliance on marginalized communities to moderate themselves requires substantial labor, especially for trans tech creators and their communities. However, larger platforms can learn from these smaller scale implementations to better protect and moderate content for trans communities and marginalized people broadly. Our work extends prior research on trans technologies [27], content moderation [26], and community-based approaches [49, 54] by contributing a framework for trans-centered moderation and considering how trans tech creators built and employed moderation systems that better serve trans people. Both

[5]Jailbreak the Binary is a Chrome extension that removes gender from pronouns and other terms. Gender Neutralize is a Chrome extension that turns "unnecessarily gendered words" into gender-neutral terms.

trans-centered moderation and the moderation systems discussed here can speak not only to trans contexts, but marginalized contexts broadly, as we discuss below.

5.1 A framework for trans-centered moderation

Trans-centered moderation values and considers trans contexts when designing and implementing content moderation processes. Trans tech creators, when asked, are often quick to point out online platforms' faults when it comes to navigating online spaces while trans. To combat this, they often must create their own moderation systems, and perform additional moderation labor that non-trans and non-marginalized tech creators do not have to. These moderation systems are designed specifically for trans experiences online, as they revolve around considering trans contexts alongside moderation decisions, and considering how best to protect both trans tech creators and users.

A framework of trans-centered moderation is exemplified by the work trans tech creators do to best serve their communities. Trans tech creators like Taylor Chiang (of TransZap) and Rosa Chapperri (of the Transverse and Aegis) showcase how content moderation systems can review and effectively vet transphobic users and content. Whether it is through the manual vetting of content for transphobic language, like with moderators of TransZap, or through a shared reporting system, like with Aegis, both trans technologies consider trans communities' online safety. The Product Designer for a queer social and community app we interviewed spoke to how trans moderators might speak better to trans contexts than cis moderators, but trans tech creators like Laura Horak (of the Transgender Media Portal) helped us to consider how shared understandings around transness are not monolithic. However, both showcase how appointing trans moderators can help to effectively moderate trans contexts.

Trans tech creators like Willow Hayward (of the Deadname Remover plug-in) spoke to how technologies can be anti-trans in their design. The Deadname Remover is one of many examples of trans technologies that help us to consider the limitations and constraints of technology for addressing social challenges. However, Plug-ins like Hayward's are band-aid solutions that speak to a broader need to consider trans contexts in design and broader issues of cisnormativity writ large. Others like Delilah D'Lune (of various trans technologies not named here) and Laur Bereznai (of the Trans Peer Network) spoke to the desire for collective governance or community models. D'Lune mentioned collective forms of decision-making and representative policies to help smaller subgroups have equal say compared to majority groups, while Bereznai mentioned a reshaping and reforming of community spaces to limit hierarchies and continually uplift and center the most marginalized. Both D'Lune and Bereznai's comments helped us to form a better idea of what a truly collective governance model might look like in theory.

Overall then, trans-centered moderation centers trans contexts by taking technologically-enabled transphobia seriously, and supporting trans tech creators and their communities to combat anti-trans actors and sentiments online.

5.2 Suggestions for the larger tech and platform economy

Trans-centered moderation helps not only trans people, but marginalized people more broadly, because it centers marginalized user contexts and can be similarly applied to non-trans marginalized contexts (and marginalized identities that often intersect with transness) by considering how these contexts inform content moderation. Approaches that center trans people, BIPOC people, women, and other marginalized groups may help to improve online spaces by reducing online abuse and providing more contextual moderation.

If viewed uncritically, trans-centered moderation may be seen as a technological-determinist view of technology and platforms; this means giving technology undue power in its ability to fix social ills [14, 18]. Technology cannot single-handedly alleviate issues such as transphobia or marginalization, but designing better technology and better platforms for trans users is one part of the process. Issues of transphobia and marginalization are culturally-ingrained in our society, so they require complex and long-form solutions that technology and platform design are only one part of. We suggest that to start to help trans and marginalized users broadly who interact with mainstream technologies and platforms, designers and policy managers should consider the models and approaches trans tech creators have put forth, as we have described in this paper. Trans tech creators' content moderation processes, and the tools and governance structures they design to implement them, can better inform mainstream technology's moderation processes, by taking transphobia seriously and working to counteract it. We detail the four facets of trans-centered moderation and how they might be applied to larger platforms for both trans and broadly marginalized contexts below.

5.2.1 Reviewing and vetting users and content. With concerns around safety for marginalized groups on platforms with "community-driven" [49] or bottom-up moderation approaches [6], it makes sense why trans tech creators in our study had more intensive reviewing and vetting systems than mainstream platforms. Mainstream platforms that employ algorithmic content moderation tools may not be able to apply these revisions to their systems without retraining their algorithms to better catch anti-trans or other bigoted sentiments. Platforms like Reddit and Discord, however, may implement tools to assist smaller communities in using more concerted reviewing and vetting systems. Unfortunately, more intensive vetting processes encumber already unpaid volunteer moderators with additional labor. Platforms could help by providing vetting guidance that draws from updated hate speech detection mechanisms that help identify anti-trans or otherwise harmful bad actors on their platform.

5.2.2 Appointing moderators to better consider identity-based contexts. Echoing concerns above, adding extra labor for unpaid volunteer moderators is not ideal. Platforms then should consider hiring more moderators from marginalized backgrounds to have a larger and more diverse moderator-base to pull from when moderating identity-based issues. To attempt to combat the issue of transferring bias in algorithm training [3], platforms should re-train existing algorithmic tools with the involvement of developers from marginalized backgrounds. Yet as Laura Horak (of the Transgender

Media Portal) reminded us, trans people are not all the same, and will have different ideas regarding desired moderation outcomes. Relatedly, marginalized people are not all the same, but issues they face are often shared or common across their identity group. Hiring multiple moderators from marginalized backgrounds with variation within the identity groups they belong to will help platforms better consider identity-based contexts overall. That said, given the often distressing contexts under which content moderators work [42], placing more trans people and other marginalized people in those working conditions may also be harmful.

5.2.3 Acknowledging technology's limitations for addressing social issues. Platforms want to appear as neutral spaces for free expression and discussion [22], but research shows that certain groups are disadvantaged online or disproportionately moderated more than others [26, 47]. While platforms can address this by implementing policy changes and tools recommended by scholars and platform users, platforms cannot singlehandedly solve social issues such as racism, sexism, transphobia, or ableism, amongst others. Platforms should listen to their users and consider identity-based contexts rather than employing "one-size-fit-all" solutions [48]. However, platforms must approach these issues with the knowledge that online spaces reify and reiterate structures of oppression and marginalization, rather than free users of them [37, 56]. By keeping this in mind, platforms will be better prepared to help implement changes that can benefit trans people and other marginalized users.

5.2.4 Employing collective governance and community models. While many mainstream platforms like Facebook, Instagram, and Twitter/X are run by centralized actors/stakeholders, these platforms can still learn from trans tech creators' collective governance and community models. Thach [53] argued that there is an existing feedback loop between prevalent streamers on Twitch and Twitch's administration, where influential users call out the platform and ask their communities to assist in getting the platform to change policies and listen to its users; this feedback loop could be implemented on mainstream platforms by encouraging more user feedback from smaller communities. If a platform promises its users safety and wellbeing and continually works towards improving their spaces, its users can also continually work towards this as well, as both the platform and its users keep each other accountable. There are obvious power differentials between a platform's designers and policy-makers and the everyday user, but systems can be put in place to give users and their communities more power and more say into how a platform is governed. This balance is tricky but a fruitful space to explore, as too much power in designers and policy-makers' hands can result in disproportionate moderation [26, 47], but too little overhead review from platforms can result in abusive and bigoted content and the formation of extremist spaces [30, 49].

5.3 Limitations and future research

While our sample is diverse in terms of gender, we acknowledge that the vast majority live in the US and speak English. Additionally, more than 75% of participants were white. Especially given that many designers relied on their own experiences to create technologies, this poses an issue for understanding the practices and needs of those who are multiply marginalized, particularly trans people of color. Also, while the study uplifts trans tech creators and their reflections on how their design processes unfolded, it does not account for community members' or users' experiences. We encourage future studies to explore trans technologies alongside other axes of identity such as race or disability, and to explore similar questions from user perspectives. Finally, our study was conducted in a largely Western context, so our results may as easily apply to non-Western contexts. However, our findings still speak to the larger online systems that propagate anti-trans sentiments and fail to address trans users' needs online. Future studies should analyze trans tech creators' experiences in non-Western contexts and consider how best to design for both trans and generally marginalized audiences globally. Future research could further explore the interconnectedness of many smaller communities through trans governance systems such as Aegis and how this might address the issue of larger-scale implementation in context-centered content moderation approaches. Additionally, future research could consider how trans-centered moderation may be extended to apply to other types of marginalized users, but should consider that trans-specific implementations may not work across the board. Our method of learning directly from trans tech creators may work similarly for technology creators from other marginalized backgrounds.

6 CONCLUSION

Trans people have been at the forefront of various technologies across history; however, as many scholars note, these histories have often been modified or erased [10, 39], ignoring trans people's important contributions in designing technologies and online infrastructures. Although trans people and technology design, even when combined, cannot fix society's ills or mainstream online platforms, trans tech creators offer important insights into how to better design moderation systems to protect trans people. By centering transness and offering a framework of trans-centered moderation, we propose design suggestions for online platforms and communities.

At the heart of these design suggestions is an approach of care towards trans people, whether a platform is explicitly/specifically trans or includes trans users. To reduce harm towards a marginalized group, technology designers must center that marginalized group in their design and moderation processes, considering how previous design and moderation frameworks may have been exclusionary or failed to account for that group's lived experiences. As the trans tech creators in this study described, online platforms are often rife with transphobia, both from platform users and as programmed into platform infrastructures. Transphobic abuse and disproportionate content moderation underscore online experiences for trans people, and trans tech creators show us ways trans communities are already trying to account for these issues and actively work against them. We draw from their approaches to governing their own communities to contribute the concept of trans-centered moderation, which can help start to address transphobia online and also provide an example for how to listen to marginalized groups and make changes in response to their needs.

ACKNOWLEDGMENTS

We would like to thank the National Science Foundation (NSF) for funding this work under grants #1942125 and #2210841. We would also like to thank the members of the CRITLab for their feedback throughout the writing process, as well as the many wonderful trans technology creators we interviewed. Thanks as well to our ACM FAccT reviewers.

REFERENCES

[1] Laima Augustaitis, Leland A. Merrill, Kristi E Gamarel, and Oliver L. Haimson. 2021. Online Transgender Health Information Seeking: Facilitators, Barriers, and Future Directions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445091

[2] Chris Barcelos. 2022. The Affective Politics of Care in Trans Crowdfunding. *TSQ: Transgender Studies Quarterly* 9, 1 (Feb. 2022), 28–43. https://doi.org/10.1215/23289252-9475495

[3] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. In *Social Informatics (Lecture Notes in Computer Science)*, Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri (Eds.). Springer International Publishing, 405–415.

[4] Oversight Board. 2023. Gender identity and nudity decision. https://oversightboard.com/attachment/853018979320399/

[5] Oversight Board. 2023. Oversight Board overturns Meta's original decisions in the "Gender identity and nudity" cases | Oversight Board. https://www.oversightboard.com/news/1214820616135890-oversight-board-overturns-meta-s-original-decisions-in-the-gender-identity-and-nudity-cases/

[6] Ben Bradford, Florian Grisel, Tracey L. Meares, Emily Owens, Baron L. Pineda, Jacob N. Shapiro, Tom R. Tyler, and Danieli Evans Peterman. 2019. *Report of the Facebook Data Transparency Advisory Group.* Technical Report. Yale Law School, The Justice Collaboratory.

[7] Justin Buss, Hayden Le, and Oliver L Haimson. 2022. Transgender identity management across social media platforms. *Media, Culture & Society* 44, 1 (Jan. 2022), 22–38. https://doi.org/10.1177/01634437211027106

[8] Jie Cai and Donghee Yvette Wohn. 2019. Categorizing Live Streaming Moderation Tools: An Analysis of Twitch. *International Journal of Interactive Communication Systems and Technologies (IJICST)* 9, 2 (July 2019), 36–50. https://doi.org/10.4018/IJICST.2019070103 ISBN: 9782019070106 Publisher: IGI Global.

[9] Robyn Caplan. 2018. *Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches.* Technical Report. Data&Society. https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf

[10] Lynn Conway. 2018. The Disappeared: Beyond Winning and Losing. *Computer* 51, 10 (Oct. 2018), 66–73. https://doi.org/10.1109/MC.2018.3971344

[11] Juliet Corbin and Anselm Strauss. 2008. *Basics of Qualitative Research (3rd ed.): Techniques and Procedures for Developing Grounded Theory.* SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States. https://doi.org/10.4135/9781452230153

[12] Christina Criddle. 2020. Transgender users accuse TikTok of censorship. https://www.bbc.com/news/technology-51474114

[13] Avery Dame-Griff. 2023. *The two revolutions: a history of the transgender internet.* New York University Press, New York.

[14] Marc J. De Vries. 2017. Philosophy as Critique. In *Critique in Design and Technology Education*, P John Williams and Kay Stables (Eds.). Springer Singapore, Singapore, 15–30. https://doi.org/10.1007/978-981-10-3106-9_2 Series Title: Contemporary Issues in Technology Education.

[15] Daniel Delmonaco, Samuel Mayworm, Hibby Thach, Josh Guberman, A Augusta, and Oliver L Haimson. 2024. "What are you doing, TikTok?": How social media users perceive, theorize, and "prove" shadowbanning. In *Proceedings of the ACM on Human-Computer Interaction, 38.*

[16] Michael Ann DeVito. 2022. How Transfeminine TikTok Creators Navigate the Algorithmic Trap of Visibility Via Folk Theorization. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 380:1–380:31. https://doi.org/10.1145/3555105

[17] Christina Dinar. 2021. The state of content moderation for the LGBTIQA+ community and the role of the EU Digital Services Act. (2021).

[18] Val Dusek. 2006. *Philosophy of technology: an introduction.* Blackwell Pub, Malden, MA ; Oxford. OCLC: ocm61461614.

[19] Niki Fritz and Amy Gonzalez. 2018. Not the Normal Trans Story: Negotiating Trans Narratives While Crowdfunding at the Margins. (2018).

[20] A Gagliardi. 2023. Does Instagram Shadowban Accounts? https://later.com/blog/instagram-shadowban/

[21] Sarah Gilbert. 2023. Towards Intersectional Moderation: An Alternative Model of Moderation Built on Care and Power. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Oct. 2023), 256:1–256:32. https://doi.org/10.1145/3610047

[22] Tarleton Gillespie. 2018. *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press, New Haven. OCLC: on1005113962.

[23] GLAAD. 2023. Social Media Safety Index 2023. https://assets.glaad.org/m/7adb1180448da194/original/Social-Media-Safety-Index-2023.pdf

[24] Oliver Haimson. 2018. Social Media as Social Transition Machinery. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 63:1–63:21. https://doi.org/10.1145/3274332

[25] Oliver L. Haimson, Avery Dame-Griff, Elias Capello, and Zahari Richter. 2021. Tumblr was a trans technology: the meaning, importance, history, and future of trans technologies. *Feminist Media Studies* 21, 3 (April 2021), 345–361. https://doi.org/10.1080/14680777.2019.1678505

[26] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 466:1–466:35. https://doi.org/10.1145/3479610

[27] Oliver L. Haimson, Dykee Gorrell, Denny L. Starks, and Zu Weinger. 2020. Designing Trans Technology: Defining Challenges and Envisioning Community-Centered Solutions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376669

[28] Oliver L. Haimson and Anna Lauren Hoffmann. 2016. Constructing and enforcing "authentic" identity online: Facebook, real names, and non-normative identities. *First Monday* (June 2016). https://doi.org/10.5210/fm.v21i6.6791

[29] Oliver L. Haimson, Kai Nham, Hibby Thach, and Aloe DeGuia. 2023. How Transgender People and Communities Were Involved in Trans Technology Design Processes. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–16. https://doi.org/10.1145/3544548.3580972

[30] Daniel G Heslep and PS Berge. 2021. Mapping Discord's darkside: Distributed hate networks on Disboard. *New Media & Society* (Dec. 2021), 14614448211062548. https://doi.org/10.1177/14614448211062548 Publisher: SAGE Publications.

[31] Anna Lauren Hoffmann and Anne Jonas. 2016. Recasting Justice for Internet and Online Industry Research Ethics. https://papers.ssrn.com/abstract=2836690

[32] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 150:1–150:27. https://doi.org/10.1145/3359252

[33] Sophia Cecelia Leveque and Gwendolyn Ann Smith. 2017. *Trans / Active: A Biography of Gwendolyn Ann Smith.* Library Partners Press.

[34] Adrienne Massanari. 2017. #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society* 19, 3 (March 2017), 329–346. https://doi.org/10.1177/1461444815608807

[35] Joseph Alex Maxwell. 2013. *Qualitative research design: an interactive approach* (3rd ed ed.). Number v. 41 in Applied social research methods. SAGE Publications, Thousand Oaks, Calif.

[36] Samuel Mayworm, Michael Ann DeVito, Dan Delmonaco, Hibby Thach, and Oliver L. Haimson. 2023. Content Moderation Folk Theories And Perceptions of Platform Spirit Among Marginalized Social Media Users. *ACM Transactions on Social Computing* (Dec. 2023), 3632741. https://doi.org/10.1145/3632741

[37] Lisa Nakamura. 2007. *Digitizing race: visual cultures of the Internet.* Number 23 in Electronic mediations. University of Minnesota Press, Minneapolis. OCLC: ocn154789875.

[38] Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "I'm fully who I am": Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1246–1266. https://doi.org/10.1145/3593013.3594078

[39] Whitney (Whit) Pow. 2021. A Trans Historiography of Glitches and Errors. *Feminist Media Histories* 7, 1 (Jan. 2021), 197–230. https://doi.org/10.1525/fmh.2021.7.1.197

[40] Organizers Of Queerinai, Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J. Sutherland, Davide Locatelli, Eva Breznik, Filip Klubicka, Hang Yuan, Hetvi J, Huan Zhang, Jaidev Shriram, Kruno Lehman, Luca Soldaini, Maarten Sap, Marc Peter Deisenroth, Maria Leonor Pacheco, Maria Ryskina, Martin Mundt, Milind Agarwal, Nyx Mclean, Pan Xu, A Pranav, Raj Korpan, Ruchira Ray, Sarah Mathew, Sarthak Arora, St John, Tanvi Anand, Vishakha Agrawal, William Agnew, Yanan Long, Zijie J. Wang, Zeerak Talat, Avijit Ghosh, Nathaniel Dennler, Michael Noseworthy, Sharvani Jha, Emi Baylor, Aditya Joshi, Natalia Y. Bilenko, Andrew Mcnamara, Raphael Gontijo-Lopes, Alex Markham, Evyn Dong, Jackie Kay, Manu Saraswat, Nikhil Vytla, and Luke Stark. 2023. Queer In AI: A Case Study in Community-Led Participatory AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT*

'23). Association for Computing Machinery, New York, NY, USA, 1882–1895. https://doi.org/10.1145/3593013.3594134

[41] Reddit. 2022. Moderation Tools. https://support.reddithelp.com/hc/en-us/sections/15483198381332

[42] Sarah T. Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media.* Yale University Press, New Haven, CT.

[43] Julian A. Rodriguez. 2023. LGBTQ Incorporated: YouTube and the Management of Diversity. *Journal of Homosexuality* 70, 9 (July 2023), 1807–1828. https://doi.org/10.1080/00918369.2022.2042664

[44] Mey Rude. 2019. Trace Lysette Is Latest Trans Woman Banned By Tinder. https://www.out.com/transgender/2019/9/19/trace-lysette-latest-trans-woman-be-banned-tinder

[45] Salty. 2019. *Exclusive: An Investigation into Algorithmic Bias in Content Policing on Instagram.* Technical Report. Salty. https://saltyworld.net/algorithmicbiasreport-2/ Section: #MeToo.

[46] Laura Savolainen. 2022. The shadow banning controversy: perceived governance and algorithmic folklore. *Media, Culture & Society* 44, 6 (Sept. 2022), 1091–1109. https://doi.org/10.1177/01634437221077174

[47] Morgan Klaus Scheuerman, Stacy M. Branham, and Foad Hamidi. 2018. Safe Spaces and Safe Places: Unpacking Technology-Mediated Experiences of Safety and Harm with Transgender People. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–27. https://doi.org/10.1145/3274424

[48] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. 2021. Drawing from justice theories to support targets of online harassment. *New Media & Society* 23, 5 (May 2021), 1278–1300. https://doi.org/10.1177/1461444820913122 Publisher: SAGE Publications.

[49] Joseph Seering. 2020. Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 107:1–107:28. https://doi.org/10.1145/3415178

[50] Ellen Selkie, Victoria Adkins, Ellie Masters, Anita Bajpai, and Daniel Shumer. 2020. Transgender Adolescents' Uses of Social Media for Social Support. *Journal of Adolescent Health* 66, 3 (March 2020), 275–280. https://doi.org/10.1016/j.jadohealth.2019.08.011

[51] S Smith, Oliver L Haimson, C Fitzsimmons, and N Echarte Brown. 2021. *Censorship of Marginalized Communities on Instagram.* Technical Report. Salty. https://saltyworld.net/product/exclusive-report-censorship-of-marginalized-communities-on-instagram-2021-pdf-download/

[52] Denny L. Starks, Tawanna Dillahunt, and Oliver L. Haimson. 2019. Designing Technology to Support Safety for Transgender Women & Non-Binary People of Color. In *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*. ACM, San Diego CA USA, 289–294. https://doi.org/10.1145/3301019.3323898

[53] Hibby Thach. 2023. *"Should We Unban, Chat?": Communal Content Moderation on Twitch.* Ph. D. Dissertation. University of Illinois at Chicago, Chicago, IL. https://www.proquest.com/openview/fdab77b69ba028f30f6315714c7ec144/1?pq-origsite=gscholar&cbl=18750&diss=y

[54] Hibby Thach, Samuel Mayworm, Daniel Delmonaco, and Oliver Haimson. 2022. (In)visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit. *New Media & Society* (July 2022), 14614448221109804. https://doi.org/10.1177/14614448221109804

[55] Tumblr Staff. 2018. A better, more positive Tumblr. https://staff.tumblr.com/post/180758987165/a-better-more-positive-tumblr

[56] Jacqueline Ryan Vickery. 2018. This Isn't New: Gender, Publics, and the Internet. In *Mediating Misogyny: Gender, Technology, and Harassment*. Springer International Publishing, Cham, 31–49. https://doi.org/10.1007/978-3-319-72917-6_2

[57] Qunfang Wu and Bryan Semaan. 2023. "How Do You Quantify How Racist Something Is?": Color-Blind Moderation in Decentralized Governance. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–27. https://doi.org/10.1145/3610030

Received 22 January 2024; revised 30 April 2024; accepted 30 April 2024