Misgendered During Moderation: How Transgender Bodies Make Visible Cisnormative Content Moderation Policies and Enforcement in a Meta Oversight Board Case

Samuel Mayworm
University of Michigan
Ann Arbor, Michigan, USA

Kendra Albert
Harvard University
Cambridge, Massachusetts, USA

Oliver L. Haimson
University of Michigan
Ann Arbor, Michigan, USA

ABSTRACT

Transgender and nonbinary social media users experience disproportionate content removals on social media platforms, even when content does not violate platforms' guidelines. In 2022, the Oversight Board, which oversees Meta platforms' content moderation decisions, invited public feedback on Instagram's removal of two trans users' posts featuring their bare chests, introducing a unique opportunity to hear trans users' feedback on how nudity and sexual activity policies impacted them. We conducted a qualitative analysis of 83 comments made public during the Oversight Board's public comment process. Commenters criticized Meta's nudity policies as enforcing a cisnormative view of gender while making it unclear how images of trans users' bodies are moderated, enabling the disproportionate removal of trans content and limiting trans users' ability to use Meta's platforms. Yet there was significant divergence among commenters about how to address cisnormative moderation. Some commenters suggested that Meta clarify nudity guidelines, while others suggested that Meta overhaul them entirely, removing gendered distinctions or fundamentally reconfiguring the platform's relationship to sexual content. We then discuss how the Oversight Board's public comment process demonstrates the value of incorporating trans people's feedback while developing policies related to gender and nudity, while arguing that Meta must go beyond only revising policy language by reevaluating how cisnormative values are encoded in all aspects of its content moderation systems.

CCS CONCEPTS

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

KEYWORDS

algorithmic content moderation, cisnormativity, content moderation, Meta, nonbinary, nudity, Oversight Board, social media, transgender

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

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658907

ACM Reference Format:
Samuel Mayworm, Kendra Albert, and Oliver L. Haimson. 2024. Misgendered During Moderation: How Transgender Bodies Make Visible Cisnormative Content Moderation Policies and Enforcement in a Meta Oversight Board Case. 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, 12 pages. https://doi.org/10.1145/3630106.3658907

1 INTRODUCTION

> *"When we submitted our case, we didn't imagine this would be the result or that anyone would even care – but we did it anyways. So, in other words, always speak up even if you think nobody is listening... you might just f around and free the nip."* – @heycolanda, 2023[35]

In 2021 and 2022, Meta (Instagram and Facebook's parent company) removed two trans[1] individuals' top surgery fundraising posts from Instagram (see Figure 1). The posts contained bare-chested images of the two users with their nipples covered; the posts were removed for allegedly violating Instagram's Adult Nudity and Sexual Activity Policy and Sexual Solicitation Policy, despite the posts not containing sexual solicitation or nudity that violates Meta's guidelines.

Figure 1: The two removed Instagram photos (@heycolanda)

The two users appealed Instagram's removal of their posts to the Oversight Board, an independent organization that Meta created to oversee content moderation decisions on its platforms, and the Board accepted their appeal case for review. The Oversight Board opened the appeal case to public comment as part of their review

[1]We use the term "trans" throughout this work to describe a broad range of transgender people, explicitly including nonbinary trans people and their experiences. The two Instagram users described throughout this work are trans and nonbinary.

process, encouraging interested commenters to provide feedback on whether Meta's content moderation policies and practices "sufficiently respect the rights of trans and nonbinary users." [55]. After completing the appeal evaluation, including reviewing 130 public comments related to the two Instagram users' appeal case, the Oversight Board overturned Meta's removal of the posts, ruling that Meta had incorrectly applied their Adult Nudity and Sexual Activity and Sexual Solicitation Guidelines. The Oversight Board publicly published 84 of the 130 public comments they received as part of their official appeal decision, excluding comments that either included identifying details about people other than the commenter, were "clearly irrelevant, abusive, or disrespectful of the human and fundamental rights of any person or group of persons" [55]," or when commenters requested not to publicly publish their comments. These comments provide a rare and unique window into the relationship between trans users and online platforms, as commenters described how they perceived Meta's policies related to nudity, and its content moderation practices more broadly, as enabling disproportionate removal of trans users' content and accounts.

Though most social media platforms forbid many types of nude or sexual content, many platforms exempt certain categories of nude content from removal, such as nudity in healthcare imagery. The two trans users' posts contained imagery related to gender affirming surgery, which was already explicitly allowed in Meta's Adult Nudity and Sexual Activity policy [19]. However, the two trans Instagram users experienced removal for "sexual solicitation," "seemingly because [their posts] include breasts and a link to a fundraising page" [54], even though the images in their posts were explicitly related to top surgery (a form of gender affirming healthcare, thus explicitly exempted in Meta's own policies).

In general, trans social media users and their content are disproportionately removed from social media platforms for "adult" or "explicit" content, even when their content does not violate platforms' policies or does not contain nudity in the first place [32, 63], leading some trans users to theorize that social media platforms' policies and content moderation practices impose cisnormative values on trans users [45]. Cisnormativity is defined as "the assumption that all people are cisgender, i.e. [having] a gender identity and presentation that are consistent with the sex they were assigned at birth" [14], while transgender people and experiences are either unaccounted for or are considered "abnormal" [68]. Past literature identifies how cisnormative values are embedded in sociotechnical systems such as social media sites [5, 6], including social media content moderation systems that evaluate people and their bodies – and how those systems may exclude or perpetuate discrimination against trans people [14, 32, 41, 48, 57]. Past work also describes how social media content moderation systems exclude trans users from consideration in their platforms' policies and while designing policy enforcement tools, enabling the disproportionate incorrect removals of trans users and their content [32]. The two trans Instagram users' appeal case, along with the public comments submitted to the Oversight Board, highlights how cisnormative content moderation systems enable trans users' exclusion on social media platforms.

In this paper, we examine public comments related to the two trans Instagram users' appeal to the Oversight Board, predominantly written by trans people or representatives of LGBTQIA+ advocacy organizations. We ask:

- **RQ1:** In what ways might trans social media users perceive Instagram's policies as imposing cisnormative values related to bodies and nudity on trans users and their content?
- **RQ2:** How do trans social media users recommend platforms address cisnormative values embedded in their policies and content moderation systems?

To address our research questions, we qualitatively analyzed 83 of the 84 public comments[2] published by the Oversight Board related to the two trans Instagram users' appeal. We found that the public commenters overwhelmingly perceived Meta's Adult Nudity and Sexual Activity guidelines to exclude trans users, imposing unclear content moderation policies and procedures along with cisnormative views of bodies and gender presentation. However, users were split on how Meta should address their policies related to nudity and sexual content, with some commenters suggesting that Meta clarify their nudity policy language or replace gendered policy language with gender-neutral terminology, and others suggesting that Meta stop moderating nudity differently based on gender. We then discuss how the Oversight Board's use of trans people's policy insights to guide their appeal decision and policy recommendations to Meta is an example of effectively involving trans people in social media policy decisions relating to gender. We argue that social media corporations like Meta must seek feedback from trans policy experts while revising gender and nudity policies that disproportionately impact trans users; ideally, they should preemptively work with trans policy experts while initially developing platforms' policies relating to gender. This study contributes to the literature by specifically analyzing public commenters' perspectives on how cisnormative values are embedded in Meta's content moderation systems and enable the disproportionate removal of content featuring trans bodies, along with public commenters' differing opinions on how Meta should address cisnormative moderation on its platforms.

2 BACKGROUND

2.1 Meta's Adult Nudity and Sexual Activity Policy and Sexual Solicitation Policy

Meta's Adult Nudity and Sexual Activity Policy regulates nude or sexual content on their platforms, including Instagram [19]. The policy describes which content is forbidden, including imagery of nude adults featuring "female nipples" [19]. The policy allows certain nude content if accompanied with a sensitivity label, including nudity related to gender-affirming surgery [19]. Meta also allows exceptions to the "female nipple" ban in certain medical contexts (including gender-confirming surgery contexts) [19]. Meta refers to "female breasts" and "female nipples" to describe specific kinds of chest nudity, but the policy does not explicitly define these terms. The only explicitly trans-related language in Meta's nudity policy

[2]One of the 84 public comments was submitted jointly by two authors of this paper; this comment was excluded from our analysis.

relates to their gender-confirming surgery exemption and sensitivity warning requirement; Meta's nudity policy language does not explicitly address how trans bodies, including chest nudity, should be moderated.

Additionally, Meta's Sexual Solicitation Policy, separate from its Adult Nudity and Sexual Activity policy, forbids content that "facilitates, encourages, or coordinates sexual encounters or commercial sexual services between adults;" the explicit solicitation of funds in exchange for "nude photos [or] imagery" creates a potential policy violation [20]. The two trans Instagram users' posts were removed for violating Meta's Sexual Solicitation Policy, likely "because they contain breasts and a link to a fundraising page" [54]. As Meta's Adult Nudity and Sexual Activity Policy contains no specific guidance for moderating trans nudity, the posts were likely removed for containing "female" chest nudity – despite neither user being female, and neither post showing exposed nipples. Likewise, the posts were identified as violating Meta's Sexual Solicitation Policy despite the fundraiser raising money for top surgery funds and not for sexual activity. These instances of content removal highlight how Meta's nudity and sexual solicitation policies do not account for trans users and their bodies, making it unclear how their content will be moderated.

2.2 The Oversight Board

The Oversight Board is an independent, Meta-funded body of 22 policy, social justice, and technology experts who oversee content moderation policy and decisions on Meta's platforms [50, 52, 53]. Users can appeal to the Oversight Board regarding instances of content or account removal. The Board does not review every appeal request, but instead chooses key "highly emblematic" appeal cases to determine whether Meta's moderation decisions align with their platform policies, prioritizing "cases that are challenging, globally relevant, and can inform future policy" [54]. Although the Oversight Board emphasizes its independence and willingness to overturn Meta's content moderation decisions, critics argue that the Board's existence could potentially "whitewash" Meta's past history of unethical policies and harmful platform content [2, 59], and question the Board's value given that Meta is not required to follow its recommendations.

The two trans Instagram users filed an appeal to the Oversight Board after Meta removed their posts for allegedly violating its Sexual Solicitation Policy; the appeal case was one of the "highly emblematic" cases selected for review. The Oversight Board's appeal cases are open to public comment, garnering public feedback, opinions, and socio-political context related to the case. The Oversight Board received 130 comments regarding the two trans Instagram users' appeal, publishing 83 on their website. Following careful review, the Oversight Board either upholds or overturns Meta's moderation decisions, and provides a public, written explanation of the Board's reasoning. Explanations may also include explicit recommendations for modifying Meta's policies beyond upholding or reversing their moderation decision. In this case, the Oversight Board overturned Meta's decision to remove the two trans Instagram users' posts, publishing an explanation for their decision along with policy recommendations beyond overturning the removal. The Oversight Board had previously reviewed similar nudity-related appeals, including overturning Instagram's removal of a breast cancer awareness post that included images of "visible and uncovered female nipples" [49]. However, the two trans Instagram users' appeal was the first Oversight Board case to specifically address how trans users and their bodies are moderated under Meta's nudity guidelines.

3 LITERATURE REVIEW

We provide an overview of prior work related to content moderation and nudity, along with prior work related to the disproportionate moderation experiences of trans users on social media platforms.

3.1 Content Moderation and Nudity

Social media platforms often employ content moderation to remove nude or sexual imagery from their platforms [24, 25, 31, 58, 60]. Many platforms exempt certain kinds of nude content from their nudity policies, such as in art or educational content [19, 25, 56]. However, most platforms ban or heavily restrict explicitly sexual nude content [25]. Human content moderators cannot realistically review all potentially nude imagery posted on platforms [1, 10, 39], so many platforms rely on algorithmic moderation systems as more scalable solutions for moderating rule-breaking nude imagery [9, 38, 39, 43]. However, even where human and algorithmic content moderation are both employed, platforms' moderation systems may struggle with "grey-area" nude content that is either not explicitly addressed within the platforms' guidelines or blurs the boundaries between "appropriate" and "inappropriate" nudity [18, 25, 32] – boundaries which are always culturally contingent [25].

Past work indicates that platforms' moderation of potentially nude or sexual content disproportionately affects marginalized social media users [32, 63], with permissible or "grey area" content disproportionately flagged as "nude" or "sexually explicit" when posted by women [23] and LGBTQIA+ users [11, 42, 70], particularly trans users [32]. Blunt & Stardust [7] also describe how "whore stigma" embedded in platforms' policies enable the policing and deplatforming of sex workers, reducing their access to online safety resources while impacting their ability to advertise, access digital payments, or participate in mutual aid. Gerrard & Thornham [23] criticize content moderation systems and practices as sexist, as they "impose rigid gender roles" on social media users, typically (though not exclusively) on women.

3.2 Content Moderation and Trans Social Media Users

Trans people uniquely rely on social media to meet their trans-specific needs, such as seeking out trans healthcare information [3], crowdfunding for gender-affirming healthcare [4, 21], visibility and activism [12], expressing trans identity [16], and finding community among other trans people [8, 29, 30, 37, 64, 67]. However, past work has found that trans social media users experience inequitable moderation on social media platforms, resulting in the disproportionate removal of their accounts and content even when they have not violated platforms' guidelines [15–17, 27, 31–33, 63, 69]. Inequitable content moderation concerns both content featuring trans bodies [17, 32, 63] and content related to trans activism [16]. Trans social media users are also disproportionately likely to experience

the deliberate incorrect reporting of their content by other social media users, exacerbating unwarranted removals [26, 27, 36, 62]. Although social media platforms use content moderation systems to remove illegal and harmful content [22, 25], these systems are frequently designed and function in a way that enables discriminatory moderation against trans social media users [17, 27, 32, 47].

Platforms such as YouTube [61] and TikTok [16] have faced criticism for algorithmically suppressing content posted by trans users, resulting in some trans people fighting back; for example, transfeminine TikTok users often adjust their posting behavior to prevent such suppression on the platform [16]. Platforms' "overblocking" of trans content [46] can frustrate trans users and curtail their social media use in important, trans-specific ways [17, 32, 63]. Ultimately, unequal content moderation imposes cisnormative values on users by policing users' bodies and self-expression on social media platforms, preventing them from using social media overall as freely or as safely as cisgender social media users.

While prior research has examined how content moderation impacts trans people and those posting nudity, by analyzing public comments responding to a specific case when Meta removed "explicit" content posted by trans users, we provide unique insight on how cisnormative values are embedded in content moderation systems and how commenters suggest moving forward.

4 METHODS

4.1 Data Collection

To answer our research questions, we conducted qualitative analysis of 83 of the 84 public comments published by the Oversight Board as part of the appeal's public comment process, excluding one public comment submitted by two authors of this paper. As part of their announcement that Meta's removals would be overturned, The Oversight Board published 84 of the 130 submitted comments, omitting comments if either the commenters requested that their feedback not be published publicly, if the comments themselves were "clearly irrelevant, abusive, or disrespectful of the human and fundamental rights of any person or group of persons and therefore violating [the Oversight Board's] Terms for Public Comment," or if the comments included identifying information about individuals other than the commenter [55]. The Oversight Board sought out "public comments that address whether Meta's Adult Nudity and Sexual Activity policy respects the rights of trans and nonbinary users, whether the gender confirmation surgery exemption in the policy is effective in practice, whether Meta has sufficient measures in place to reduce the risk of incorrect removals, [and] how Meta's use of automated moderation to detect nudity and sexual content could be improved" [55]. The Oversight Board also sought out comments concerning "the socio-political context [and] challenges" related to trans rights, access to gender-affirming healthcare, and gender expression, and "the role of social media as resources and forums for expression for trans and nonbinary users" [55]. Commenters were given the option to either anonymize their comments or to attribute their comments publicly. In this paper, we refer to all commenters by the public comment number included in their submission. The comments represented a range of regions; of the 130 total submitted comments, 97 comments came from the United States or Canada, 19 from Europe, 10 from Asia Pacific and Oceania, and 1 each from Central & South Asia, Latin America & Caribbean, Middle East & North Africa, and Sub-Saharan Africa.

4.2 Data Analysis

We conducted qualitative open coding [13] of the 83 public comments. First, two authors both coded three of the same public comments separately, then met to discuss codes and collaboratively refine the codebook (see Table 1). The first author then coded the remaining transcripts individually; the first author used the codebook to code the remaining transcripts, and updated the codebook with new codes developed throughout subsequent analysis. Throughout the coding process, the first author and the third author discussed emerging codes regularly. Following open coding, the three authors conducted axial coding to group codes into larger categories [13]. Themes that we developed in our data analysis include: trans erasure in Meta's nudity policy and content moderation practices; clarifying how Meta's nudity policies apply to trans users; omitting gender from Meta's policies and moderation of nudity; Meta's algorithmic moderation of trans bodies; challenges in enforcing Meta's nudity policy; challenges to sharing trans healthcare content on social media platforms; the disproportionate moderation of marginalized users' content; and including trans people in policy design. This paper predominantly focuses on the first three themes.

5 RESULTS

Below, we describe our assessment of 83 public comments published by Meta related to the two trans Instagram users' appeal case. 27 public commenters explicitly stated that they are either trans themselves or represented an organization advocating for trans people. 42 public commenters suggested that they spoke from a trans experience without explicitly stating that they are trans. Trans commenters often drew feedback from their own experiences having content moderated on Meta's platforms. 78 of the 83 published public comments were trans-affirming in nature; though the Oversight Board stated that they did not publish comments that were "abusive or disrespectful of the human and fundamental rights of any person or group of persons" [55], 5 published public comments were explicitly anti-trans and did not fall under the themes we discuss in this paper. We first describe the public commenters' criticisms of Meta's policies related to nude or sexual content and Meta's enforcement of those policies, particularly criticism of cisnormative language and viewpoints embedded in Meta's nudity policies that exclude trans users while making it unclear how content featuring their bodies should be moderated. We then describe the public commenters' suggestions for how Meta could improve their nudity policies and moderation to better include trans people, such as potentially replacing gender-specific terminology in their nudity policies, stopping the differing moderation of nudity based on gender entirely, or clarifying how human moderators should enforce nudity policies.

5.1 Commenters' Critiques of Meta's Adult Nudity & Sexual Activity Policy and Enforcement

Commenters overwhelmingly criticized Meta's Adult Nudity & Sexual Activity policy as imposing a cisnormative view of gender on trans users. PC-10613 argued that Meta's nudity policy imposes *"a binary understanding of gender that ignores the existence of non-binary users"* while moderating images of trans bodies, exacerbated *"by the absence of references to trans and nonbinary identities within [the policy]."* Several commenters described Meta's policies as *"cisnormative," "restrictive,"*, and *"heteronormative;"* PC-10637 argued that *"cisgender system designers often project notions of binary sex/gender concepts onto trans people while designing platforms and policies,"* suggesting the same may have occurred while Meta's nudity policies were developed. Many commenters objected to Meta's moderation of trans bodies under cisnormative policies that omit trans users and bodies. PC-10604 criticized Meta's cisnormative policies as *"not making sense in the context of [trans] people,"* while PC-10619 argued that *"gender expression is not binary, so the rules on whether nudity is or is not allowed should not be binary either."* PC-10613 argued that the nudity policy's omission of trans users and bodies results in the cisnormative moderation that *"does not respect the rights of trans users."* Overall, these comments highlight commenters' objections to the cisnormative view of gender and nudity encoded in Meta's nudity policies, while sharing concerns that Meta's enforcement of their cisnormative nudity policies may result in incorrectly moderating trans users' content, such as the two trans Instagram users whose posts were incorrectly removed.

Several commenters specifically described how the cisnormative and binary language throughout Meta's Adult Nudity & Sexual Activity policy excludes trans users and bodies, while making it unclear how Meta's platforms moderate content featuring trans bodies. PC-10613 argued that *"the inclusion of gendered body parts within the Adult Nudity and Sexual Activity Community Standard,"* including binary terms such as "uncovered female nipples" or "male and female genitalia," can *"create uncertainty about whether nonbinary users can show their nipples."* PC-10587 asked, *"how does Meta intend to differentiate between 'female breasts' and breasts of people of other genders?"*, emphasizing that *"without putting one's gender identity in their bio (as pronouns alone... do not provide a reliable indication of gender), there is no way to tell."* PC-10587's question gives an example of how trans users are not accounted for in Meta's nudity policy, as the policy's description of certain body parts as "female" fails to acknowledge users who are not women but still have those body parts. Further, policy language that uses binary language to describe body parts does nothing to inform nonbinary users how content featuring their bodies will be moderated. PC-10544 argued that the cisnormative, binary descriptions of body parts in Meta's nudity policies could result in *"the policing of non-heteronormative bodies."* PC-10616 similarly stated that the policy's cisnormative language fails to acknowledge that *"not all nipples are gendered,"* or that they may be *"gendered in a way that is unintelligible to cis-normative assumptions,"* resulting in a policy that *"clearly discriminates in its differentiation of whose nipples are considered 'nude'."* In the case of the two trans Instagram users, removing images containing their bare chests enforced the policy's cisnormative, binary interpretation of their chests as "female," functionally misgendering the two users while inaccurately moderating their content.

Several commenters also questioned whether Meta's algorithmic moderation systems can correctly enforce gender-specific nudity policies, let alone enforce them fairly in the context of trans users and their content. For example, PC-10628 argued that *"Meta's automated moderation systems may struggle to accurately identify the gender of trans and non-binary users, rendering [the Adult Nudity and Sexual Activity policy's] gender-confirming surgery exemption ineffective in practice."* PC-10604 expressed concern that algorithmic misassessment of gender may result in trans users being algorithmically misgendered by the platform, arguing that *"to remove [the two users'] posts as a result of exposed nipples is to say that the creators are basically 'female,' despite their personal assertion that they are neither female nor male."* PC-10593 added that *"assuming the gender of an individual based on their appearance and presentation alone is likely to lead to misgendering and calling their identity into question."* Put together, the commenters' observations highlight ways in which trans bodies make visible the cisnormativity encoded in Meta's adult nudity policy and algorithmic moderation tools, resulting in trans users' bodies being incorrectly algorithmically assessed and removed for "nudity," such as with the two trans users' Instagram posts featuring images of their chests.

Many commenters stated that Meta's unclear policies related to gender and nudity, combined with Meta's inaccurate algorithmic assessment of gender and nudity, result in Meta's disproportionate removal of trans users' content, including incorrectly removing the two trans users' Instagram posts. PC-10628 argued that the *"binary gendered nature of Meta's Nudity policies... introduces confusion [and] lacks clarity on enforcement for anyone outside the gender binary,"* which *"makes it difficult [for Meta] to mitigate... the risk of mistakenly removing"* trans users' content. Several commenters described how their own posts featuring trans bodies may be disproportionately moderated under Meta's unclear nudity policies and inaccurate algorithmic assessment of gender and nudity. PC-10497, a nonbinary artist, described the difficulty of posting art featuring their body that *"[deconstructs] the gender binary... and disassociates ideas of (binary) gender with individual body parts,"* arguing that it is impossible to do so *"with policies that distinguish my body as female and police it accordingly."* PC-10550, representing ACON (an Australian LGBTQIA+ HIV advocacy organization), described how ACON's HIV prevention content centering gender-diverse populations is frequently *"targeted [for removal] due to imagery or messaging deemed to be sexual solicitation or nudity,"* despite ACON's posts *"never including nudity or sexual solicitation."* PC-10550 argued that *"Meta's current policies threaten [ACON's] ability to communicate [HIV prevention] messages in a culturally relevant way,"* forcing ACON to *"manipulate the language"* of their health messaging content to reduce the risk of incorrect removals. PC-10550 argued that Meta's Adult Nudity and Sexual Activity policy and Sexual Solicitation policy, along with Meta's incorrect algorithmic enforcement of these policies, *"continue to represent a barrier to effective and agile engagement"* with LGBTQIA+ users, particularly *"trans and gender-diverse users... [whose] bodies may not be easily categorized as male, female, or another gender by sight."* Experiences like PC-10550's and PC-10497's, along with the two

trans Instagram users, highlight how Meta's unclear policies and inaccurate algorithmic moderation disproportionately impact trans users and their content, limiting trans users' ability to use Meta's platforms for trans self-expression, sharing healthcare information, or to otherwise meet their identity-related needs.

Commenters described different ways in which trans users are uniquely impacted by their decreased visibility on Meta's platforms due to content featuring their bodies being disproportionately moderated, such as experiencing difficulty sharing content including trans bodies generally, limitations on trans community-building, and the difficulty of sharing trans-related educational or healthcare content. Some commenters discussed the emotional and mental health impact of trans users' bodies being disproportionately moderated in ways that implicitly misgender them. PC-10579 argued that *"[the two users] are not women and regulating their chests as such is not respecting their gender,"* while PC-10613 described the *"significant risk that [nonbinary users] will be misgendered during moderation, negatively impacting their mental health."* Some commenters argued that the disproportionate moderation of trans users and bodies results in decreased trans visibility on Meta's platforms, directly harming both trans users posting content featuring their bodies and trans users seeking out content featuring bodies like their own. PC-10564 described online trans visibility as *"so important,"* particularly in the context of *"celebrating their bodies... in this case, the process of [one trans Instagram user's] transition and the other's celebration of their non-binary identity."* PC-10481 elaborated on why online visibility can be uniquely high-stakes for LGBTQIA+ users, including trans users:

> *"Many LGBTQIA+ people... are isolated from real world connections... by nature of isolation [and] being unsafe in their local population. This means that we are active users of online platforms, [where] we safely connect with each other, share our culture, learn how to look after our health, share our pride, hope, and determination. This also means that Meta has a moral duty to provide their service in a safe way to this population."*

Similarly, PC-10564 addressed the importance of online trans visibility, stating that *"representation can literally save lives by letting people know they are not alone!!!"* PC-10578 similarly stated that *"trans folks need to see people who look like them, and inhibiting this is a lethal position [for Meta] to take."* Other users elaborated on specific impacts of Meta's disproportionate moderation of trans users' content featuring their bodies. PC-10508 argued that Meta's "overblocking" [46] of trans users' content limits trans community-building on their platforms, stating that *"if IG censors [trans] bodies in this way, [Meta] will cut vulnerable people off from the love and support of their community... IG is my queer home, and there are a decreasing number of places we can go to be ourselves and to be a community."* Other commenters focused on the educational value of visible trans social media content, arguing that diminishing such content's visibility impacts trans social media users' ability to find trans-related information (including healthcare information) on social media platforms. For example, PC-10588 stressed *"the importance of being able to freely share content... that focus on normalizing and informing the public on the need for gender affirming actions... in direct relation to self care, mental health and freedom."* PC-10508 stated similar sentiments, arguing that documenting gender affirming healthcare processes online *"benefits many of us in the trans and nonbinary community [by] providing us with accurate information about expectations, recovery, and general process that we cannot get anywhere else."* Overall, the commenters agreed that Meta's disproportionate moderation of trans users and their content (enabled by Meta's cisnormative nudity policy and ineffective algorithmic detection of gender) results in trans users and their bodies being erased on Meta's platforms, preventing trans people from using Meta's platforms as freely as other users or to meet their trans-specific needs.

5.2 Commenters' Suggestions for More Trans-Inclusive Content Moderation

5.2.1 Replace Gender-Specific Policy Language with Gender Neutral Language. Though most commenters overwhelmingly argued that Meta's policies and algorithmic moderation practices enforce a cisnormative view of gender on trans users, they were divided on how Meta should address this issue. Several commenters suggested that Meta should address the problem by removing gendered language from the Adult Nudity and Sexual Activity policy wherever possible. PC-10550 criticized the current cisnormative policy language that designates certain body parts as "male" or "female," suggesting that Meta revise the gendered language to *"reflect an affirming view of bodies and gender (i.e. that there are men with breasts/chest tissue, women without breasts, and non-binary people who may or may not have breasts/chest tissue)."* PC-10624 shared similar criticisms, recommending that Meta *"reconsider the use of binary terms and language to refer to bodies and anatomies."* Commenters like PC-10550 and PC-10624 argued that Meta's use of gendered policy language to describe body parts enables the cisnormative policing of trans users' bodies on their platforms, including Meta's removal of the two trans Instagram users' posts (which were removed for including "female" nudity despite neither user being women). Commenters who criticized Meta's gendered policy language did not necessarily fully oppose Meta's differing moderation of nudity based on gender. For example, despite arguing that Meta's policy language is *"disrespectful"* and imposes *"impossible"* restrictions on gender presentation for trans, nonbinary, and gender-nonconforming users, PC-10612 also argued that *"top nudity... that may be considered offensive"* should be posted behind content warnings *"to preserve traditional [societal] sensibilities regarding nudity,"* despite *"conceding that the definition of 'offensive [top nudity]' would encompass 'female-presenting nipples' for the most part."*

Instead of focusing on Meta's differing gendered moderation of nudity, these commenters argued that removing gendered descriptions of body parts from Meta's nudity policy may help prevent future incorrect removals of trans users' content, while affirming trans users' genders instead of describing their bodies in inaccurate and harmful ways. Some commenters suggested that Meta update their existing policy language with gender neutral language. PC-10550 argued that Meta's policies should *"[use] parts-based language rather than gendered language"*, such as *"saying internal or external genitals... rather than saying men's and women's genitals."* PC-10550 added that *"[Meta's] policies should refer to bodies that have [or do not have] developed breast/chest tissue... instead of referring to "male"*

or "female" bodies," arguing that gender-neutral policy language referring to chest nudity could be more inclusive of trans users. PC-10624 similarly argued that the policy's use of the phrase "female nipples" is an example of *"binary language"* imposing cisnormative ideas of gender and anatomy on trans users, resulting in posts like the two trans users' Instagram posts being incorrectly removed. Overall, commenters argued that Meta should replace gendered descriptions of body parts with gender-neutral language in their Adult Nudity and Sexual Activity policy, allowing the policy to be more inclusive of trans users while preventing future incorrect removals of content featuring trans bodies.

5.2.2 Stop Differing Moderation of Adult or Nude Content Based on Gender. Though the commenters in Section 5.2.1 argued for gender-neutral policy language related to nudity, they did not necessarily argue against moderation of nudity traditionally associated with women (such as images of breasts). However, as a step beyond just removing gendered language, many other commenters suggested that Meta completely overhaul their Adult Nudity and Sexual Activity policy to entirely avoid moderating nude or adult content based on gender. PC-10604 argued:

> *"Meta's ban on breasts is discriminatory and a double standard. I recognize that this rule is reflective of a larger social issue, but as a popular social media [company], Meta has the power to create social change and the responsibility to conduct itself fairly. Banning images of breasts contributes to the sexualization of the naked body, and disproportionately affects women and transgender people. Meta either needs to ban images of all chests or allow images of all chests."*

Multiple commenters echoed PC-10604's sentiment on allowing images of chests regardless of gender; PC-10546 stated that if *"cis men are allowed to go topless [on Meta's platforms], then so should everyone else,"* while both PC-10626 and PC-10564 argued that Meta should allow *"all nipples"* on their platforms regardless of gender. Unlike commenters in Section 5.2.1 who suggested revising the policies to use gender neutral language for specific body parts (without challenging the fact that those body parts are moderated differently), commenters like PC-10604 argued that Meta's differing moderation of nudity and body parts based on gender is inherently discriminatory and must stop entirely. PC-10604 described Meta's gender-specific moderation of bodies as *"unfairly targeting certain demographics for body parts that are not only morally neutral, but also not exclusive to one gender or sex,"* highlighting how Meta's gendered moderation of nudity prevents many groups of users, including trans users, from posting content including their bodies as freely as other users. Some commenters specifically described Meta's differing moderation of gender and nudity as imposing harmful and outdated values related to gender on its trans users. PC-10591 stated that Meta's policies impose *"outdated"* and *"sexist"* views of body parts on its users, describing the policy's framing of *"male and female nipples [as] different and breasts [as] sex organs"* as *"non-inclusive"* toward trans users. PC-10598 stated that the current Adult Nudity and Sexual Activity policy reinforces *"archaic notions"* of gender identity that do not reflect more recent *"understandings [of] gender identity as a spectrum,"* arguing that the current policy language allows *"the oppression [that] non-gender conforming people have received for centuries"* to repeat itself on Meta's platforms. Comments like PC-10598's highlight how Meta's differing moderation of gender and nudity enables the historical oppression of trans, nonbinary, and gender-nonconforming people, and that moderating nudity differently based on gender must cease entirely to prevent suppressing those historically oppressed users, including the two trans Instagram users.

5.2.3 Clarify the Adult Nudity and Sexual Activity Policy's language. Some commenters suggested that rather than adjusting or overhauling the policy, Meta should clarify the Adult Nudity and Sexual Activity policy's language to prevent excluding trans users. PC-10613 suggested that the revised policy *"provide moderators and users with clear and detailed guidelines around the Sexual Solicitation Community Standard, including examples (and clear explanations) of compliant and non-compliant photos and text."* PC-10628 added that clear explanations may assist *"[human] moderators [who] may lack the knowledge... to be trans-inclusive in their decision making,"* while suggesting that Meta update their policy to *"include definitions of sex assigned at birth, gender, and gender identity"* while *"explicitly stating which is to be used when characterizing content,"* with the goal of *"[providing] clear guidance on how to appropriately enforce this policy."* PC-10481 also recommended that Meta specifically work with *"LGBTQIA+ human rights specialists"* to clarify the policy's language and *"[to] protect trans and nonbinary people from being unfairly censored,"* suggesting that policy experts directly representing trans communities may be the best equipped to clarify Meta's policies related to gender and nudity. Comments like the above reflect users' concern that Meta's unclear nudity policies pose challenges for both trans users and human moderators, while highlighting users' belief that clarifying policy language and providing clear enforcement examples could help moderators avoid incorrectly removing trans users' content.

6 DISCUSSION

We have analyzed how users of Meta's platforms, particularly trans people, critique Meta's Adult Nudity and Sexual Activity policy and enforcement practices related to gender, bodies, and nude content. RQ1 addressed in what ways trans social media users may perceive Instagram's policies as imposing cisnormative values related to bodies and nudity on trans users and their content. We found that the public commenters critiqued Meta's gendered nudity policy language, along with Meta's algorithmic and human content moderation enforcement practices, as imposing cisnormative values related to gender and nudity on trans users while enabling disproportionate removals of those users' content and accounts. RQ2 addressed how trans social media users recommend platforms address cisnormative values embedded in their policies and content moderation systems. We found that the public commenters were divided on how Meta should revise their nudity policies and content moderation systems to be more trans-inclusive: some commenters suggested replacing gender-specific phrasing in their nudity policy or clarifying their nudity policy, while others suggested that Meta moderate nude content equivalently regardless of gender.

Drawing from our results, we next discuss ways Meta and other social media companies could prevent incorrectly removing trans users' content in the future, and avoid imposing cisnormative values

via content moderation. The Oversight Board themselves presented additional policy recommendations for Meta beyond overturning their incorrect removal of the two Instagram users' posts; these suggestions included creating "clear, rights-respecting criteria" for the Adult Nudity and Sexual Activity Standard, providing more detail on content removal criteria in the nudity policy, and revising Meta's guidance for human moderators to more accurately reflect the policy itself. Past literature has also recommended that platforms involve trans policy experts while creating moderation policy to "reduce content moderation disparities between trans and cisgender social media users" [32, 45, 63]. We expand on past literature and the Oversight Board's recommendations by presenting the Oversight Board's public comment process as an example of effectively involving trans people in policy decisions. Additionally, we argue that Meta should preemptively consult with trans policy experts to prevent incorrect content removals before major appeal disputes take place. We also argue that revising Meta's nudity policy language alone will be ineffective in protecting trans users from incorrect moderation without also reassessing - and changing - how cisnormative values are encoded in all aspects of Meta's policies and enforcement (such as in algorithmic moderation tools and human moderation practices), and how these encoded values may negatively affect trans users regardless of the policy's phrasing.

6.1 Trans Inclusion in Policy Language

Despite disagreeing on how Meta should revise their Adult Nudity and Sexual Activity policy and enforcement practices, public commenters overwhelmingly criticized the binary and cisnormative language in Meta's nudity policies, arguing that the policies' language excludes trans users while making it unclear how content featuring their bodies are moderated, thus limiting their ability to freely post and express themselves on Meta's platforms. The Oversight Board itself noted that it is "unclear" how Meta's binary policy language (such as "male or female genitalia") "is applied to people with bodies and identities that may not align with [binary] definitions" [51]. In response, many users recommended that Meta revise their gendered policy language to avoid excluding trans users. Past literature has encouraged platforms to directly involve marginalized communities while developing or revising content moderation policy [7, 32, 45]. For example, Blunt & Stardust [7] describe how sex workers are rarely included in conversations related to platforms' anti-sex work policies or legislation targeting online sex workers despite being experts on their own online exclusion. They argue that sex workers must be included in conversations about content moderation and policy, especially related to how embedded "whorephobia" in platforms' policies impacts them [7]. The Oversight Board similarly recommended that Meta "engage diverse stakeholders" while evaluating the "human rights impact" of their current nudity policy [51].

We agree that corporations like Meta should seek feedback from trans policy experts while revising policies related to gender and nudity. Indeed, the Oversight Board's use of public feedback to guide their appeal decision and policy recommendations demonstrates the effective use of trans people's policy feedback in practice. The Oversight Board gained critical, trans-specific socio-political insight from trans people's public comments, while drawing heavily from these trans perspectives while articulating their appeal decision and policy recommendations to Meta. For example, the Oversight Board directly cited PC-10624 and PC-10616 to describe how Meta's policies inform its differing moderation of women's bodies, trans bodies, and nonbinary bodies compared to its moderation of cisgender men's bodies [51]. In doing so, the Oversight Board demonstrated the effectiveness of using trans community feedback to identify policy weaknesses and embedded cisnormative ideas that uniquely harm trans users – weaknesses that likely would not have been identified by predominantly cisgender policymakers. Similar to how Blunt & Stardust argue for centering sex workers in conversations related to content moderation policies and sex work, we argue that if the Oversight Board can effectively draw from trans people's public feedback to develop policy recommendations for Meta, then Meta itself should similarly center trans policy experts' feedback while revising policies. As such, we recommend that Meta work directly with trans policy experts while revising the wording of their Adult Nudity and Sexual Activity Policy to ensure their revised policies respect trans users' rights while avoiding the cisnormative policy language that enabled trans users' incorrect content removals in the first place.

However, we also argue that social media corporations like Meta should not limit themselves to consulting with trans experts only while revising existing policy relating to gender and nudity; additionally, they should also preemptively work with those policy experts while developing platform policies relating to gender in the first place. While the Oversight Board's use of trans users' feedback was effective in forming their appeal decision and policy recommendations, trans users should not have to experience harm through incorrect content removals to initiate changes to harmful policies, nor should trans users need to contribute unpaid, time-sensitive labor to confront transphobic content moderation and policies. Preemptively working with trans policy experts while initially developing platforms' policies related to gender and nudity could prevent trans users' exclusion from platform's policies, reducing the likelihood that their content would be incorrectly removed from platforms in the first place. Not only could this approach prevent trans users from having to endure long appeal processes in hopes of restoring their incorrectly removed content, but it could also prevent platforms from needing a controversial, high-profile appeal case to identify policy flaws that could have been avoided to begin with.

6.2 Beyond Language: Challenging Cisnormative Values Embedded in Content Moderation Systems

While we agree with the Oversight Board and public commenters that Meta must revise the gendered language of their policies to avoid excluding trans users, we do not suggest that Meta's disproportionate policing of trans users (particularly content featuring their bodies) would stop after simply changing the Adult Nudity and Sexual Activity policy's vocabulary. Platforms' content moderation systems act on embedded biases while moderating marginalized users' content regardless of platform guidelines [32]. For example, past work has described how content moderation systems act

on "embedded carceral logics" [28] to over-police Black women who openly discuss racism on social media platforms [34, 44], suppressing Black women and their online speech. Similar to how content moderation systems act on embedded biases while moderating Black women's speech, we argue that without challenging the embedded cisnormative values informing Meta's initial nudity policy and enforcement practices, Meta risks continuing to forcibly impose those values on trans users regardless of how their policies are worded – enabling this marginalized group's continued erasure.

Consider PC-10612's criticism of Meta's cisnormative nudity policy in Section 5.2.1 along with their simultaneous suggestion that Meta require content warnings for "offensive" nude content to "preserve traditional sensibilities regarding nudity." Content moderation approaches like PC-10612's may challenge the gendered phrasing of platforms' nudity policies, and may even acknowledge how that gendered phrasing excludes trans users. But by only challenging Meta's nudity policy language, approaches like PC-10612's do not address how the cisnormative values that shape "traditional sensibilities regarding nudity" are embedded in Meta's content moderation practices beyond language – and thus do little to address how content moderation marginalizes gender minorities. PC-10612 themself acknowledged that their suggestion would result in Meta continuing to suppress images containing "female-presenting nipples" regardless of changes to the nudity policy's gendered phrasing. In practice, Meta would likely continue disproportionately suppressing images of trans users' bodies featuring certain kinds of chest nudity that Meta fundamentally considers "female" regardless of the policy's wording or the users' actual genders, thus delegitimizing trans users' genders while continuing to disproportionately suppress content featuring their bodies. Though the Oversight Board did not recommend PC-10612's suggested policy changes, the commenter's suggestion shows an example of what could occur if Meta only challenged the cisnormative phrasing of its nudity policies, while leaving intact the embedded cisnormative values informing content moderation practices.

Instead of only revising gendered language in nudity policies, companies like Meta should also critically reevaluate other aspects of their platforms' content moderation systems that may impose cisnormative values on trans users and result in their content being disproportionately removed. In Meta's case, one approach could include reassessing how algorithmic moderation tools enforce nudity policies, particularly how they determine gender [65, 66], whether their datasets sufficiently represent trans people and bodies [40], and whether content contains rule-breaking nudity to begin with. Meta's algorithmic moderation system incorrectly classified the two trans users' posts as nudity, and also may have gendered them as female, failing to correctly enforce Meta's policies while delegitimizing and misgendering the two users by removing their "female nipples" despite neither user being female. The incorrect algorithmic assessment and removal demonstrates how cisnormative values can be embedded in content moderation systems far beyond a policy's language. Other examples of cisnormativity embedded in content moderation likely include human moderators misinterpreting nudity policies and imposing personal cisnormative judgements of body parts and gender on trans users' bodies. In addition to reassessing algorithmic moderation systems, platforms could also assess how human moderators are trained to decide whether nude content violates platforms' nudity policies, and whether human moderators' internalized cisnormative values bias their removal decisions (a suggestion that the Oversight Board itself recommended to Meta). Failing to address how cisnormative values are embedded in content moderation systems could result in continued trans exclusion and erasure on social media regardless of how policies are worded.

6.3 Limitations

Our dataset consisted of 83 of the 130 public comments submitted to the Oversight Board, most of which were anonymous comments submitted by trans users. The Oversight Board did not publish certain comments that were either "clearly irrelevant, abusive or disrespectful of the human and fundamental rights of any person or group of persons" or included "personally identifying information regarding individuals other than the commenter;" commenters could also choose not to have their comments publicly shared. Due to the pre-set and predominantly anonymous nature of our dataset, along with the majority of comments being anonymous, we are unable to follow-up with most commenters to gather more information about their thoughts related to Meta's incorrect content removal, Adult Nudity and Sexual Activity Policy, Sexual Solicitation Policy, or overall content moderation practices. In addition, many individuals who might have relevant experiences may have ignored this call for input, assuming that it was unlikely to result in meaningful change, or not known about it. Others may not have commented due to the unpaid yet difficult and highly personal nature of writing about one's own oppression. Future research, such as interview studies with trans social media users, may glean further insight into how trans users perceive platforms' policies related to gender and nudity, how they are affected by the enforcement of those policies, and how they may recommend those policies be changed.

7 CONCLUSION

We contributed an understanding of how public commenters on a recent Oversight Board case perceive Meta's nudity policies to exclude trans users by imposing cisnormative values, enabling the disproportionate and incorrect removal of content featuring trans people's bodies. We presented commenters' differing opinions on how Meta should revise their policies and moderation practices related to gender and nudity, contrasting commenters who recommended that Meta clarify its policies or revise policies' gendered language with commenters who recommended that Meta cease moderating nudity based on gender entirely. We then discussed ways platforms like Meta could revise their Adult Nudity and Sexual Activity policy language to better include trans users, recommending that platforms work directly with trans policy experts to guide policy language revisions and to rethink policies entirely. We argue that to avoid trans exclusion and erasure online, companies like Meta must go beyond policy language updates to also critically assess and adjust how embedded cisnormative values inform content moderation and policy enforcement systems.

ACKNOWLEDGMENTS

We thank the members of the Community Research on Identity and Technology (CRIT) Lab at UMSI for their helpful feedback on our work. We also thank our anonymous reviewers for their constructive comments that improved this work. This work was supported by the National Science Foundation grant #1942125.

8 POSITIONALITY

The authors collectively represent a broad spectrum of trans, nonbinary, and queer identities and lived experiences, and are deeply familiar with (or have experienced) the identity-based harm faced by trans users on the internet. Our identities benefit our collective ability to interpret and understand the experiences faced by queer, trans, and nonbinary commenters. Two of the authors submitted public comments to the Oversight Board related to the appeal case; the two authors' comments are not cited in this paper and were excluded from the coding process.

REFERENCES

[1] Andrew Arsht and Daniel Etcovitch. 2018. The Human Cost of Online Content Moderation. https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation?onwardjourney=584162_v1

[2] Chinmayi Arun. 2021. Facebook's Faces. *SSRN Electronic Journal* (2021). https://doi.org/10.2139/ssrn.3805210

[3] 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

[4] 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

[5] Rena Bivens. 2017. The gender binary will not be deprogrammed: Ten years of coding gender on Facebook. *New Media & Society* 19, 6 (June 2017), 880–898. https://doi.org/10.1177/1461444815621527

[6] Rena Bivens and Oliver L. Haimson. 2016. Baking Gender Into Social Media Design: How Platforms Shape Categories for Users and Advertisers. *Social Media + Society* 2, 4 (Oct. 2016), 205630511667248. https://doi.org/10.1177/2056305116672486

[7] Danielle Blunt and Zahra Stardust. 2021. Automating whorephobia: sex, technology and the violence of deplatforming: An interview with Hacking//Hustling. *Porn Studies* 8, 4 (Oct. 2021), 350–366. https://doi.org/10.1080/23268743.2021.1947883

[8] 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

[9] Carlos Caetano, Sandra Avila, William Robson Schwartz, Silvio Jamil F. Guimarães, and Arnaldo De A. Araújo. 2016. A mid-level video representation based on binary descriptors: A case study for pornography detection. *Neurocomputing* 213 (Nov. 2016), 102–114. https://doi.org/10.1016/j.neucom.2016.03.099

[10] Byung Cheol Le. 2017. Inconsistent Work Performance in Automation, Can we Measure Trust in Automation? *International Robotics & Automation Journal* 3, 6 (Dec. 2017). https://doi.org/10.15406/iratj.2017.03.00075

[11] Alexander Cheves. 2018. The Dangerous Trend of LGBTQ+ Censorship on the Internet. https://www.out.com/out-exclusives/2018/12/06/dangerous-trend-lgbtq-censorship-internet

[12] Erica Ciszek, Paxton Haven, and Nneka Logan. 2023. Amplification and the limits of visibility: Complicating strategies of trans voice and representations on social media. *New Media & Society* 25, 7 (July 2023), 1605–1625. https://doi.org/10.1177/14614448211031031

[13] 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

[14] Sasha Costanza-Chock. 2018. Design Justice, A.I., and Escape from the Matrix of Domination. *Journal of Design and Science* (July 2018). https://doi.org/10.21428/96c8d426

[15] Cristina Criddle. 2020. Transgender users accuse TikTok of censorship. *BBC News* (Feb. 2020). https://www.bbc.com/news/technology-51474114

[16] Michael Ann DeVito. 2022. How Transfeminine TikTok Creators Navigate the Algorithmic Trap of Visibility Via Folk Theorization. (2022).

[17] Christina Dinar. 2021. *The state of content moderation for the LGBTIQA+ community and the role of the EU Digital Services Act.* Technical Report. Heinrich-Böll-Stiftung. 23 pages.

[18] Natasha Duarte, Emma Llansó, and Anna Loup. 2018. Mixed Messages? The Limits of Automated Social Media Content Analysis. In *2018 Conference on Fairness, Accountability, and Transparency*. https://cdt.org/files/2017/12/FAT-conference-draft-2018.pdf

[19] Facebook. 2021. Adult Nudity and Sexual Activity. https://transparency.fb.com/policies/community-standards/adult-nudity-sexual-activity/?from=https%3A%2F%2Fm.facebook.com%2Fcommunitystandards%2Fadult_nudity_sexual_activity%2F%3Fprivacy_mutation_token%3DeyJ0eXBlIjowLCJjcmVhdGlvbl90aW1lIjoxNjM3MTU2Mjc3LCJjYWxsc2l0ZV9pZCI6MTUwODA5NTI253D%26__m_async_page__%26__big_pipe_on__%26fb_dtsg_ag%3DAQwCeH0YZhbwj16xn88Fks9UHTnrTkr9xlge52JYlpiJuYGu%253A34%253A1624034963%26jazoest%3D25008

[20] Facebook. 2022. Sexual Solicitation. https://transparency.fb.com/policies/community-standards/sexual-solicitation/

[21] Niki Fritz and Amy Gonzales. 2018. Not the normal trans story: negotiating trans narratives while crowdfunding at the margins. *International Journal of Communication* 12 (2018), 20.

[22] Jason Gallo and Clare Cho. 2021. *Social Media: Misinformation and Content Moderation Issues for Congress.* Technical Report R46662. Congressional Research Service. https://crsreports.congress.gov/product/pdf/R/R46662

[23] Ysabel Gerrard and Helen Thornham. 2020. Content moderation: Social media's sexist assemblages. *New Media & Society* 22, 7 (July 2020), 1266–1286. https://doi.org/10.1177/1461444820912540 Publisher: SAGE Publications.

[24] Tarleton Gillespie. 2017. Governance of and by platforms. In *The SAGE Handbook of Social Media*. SAGE, New York, 30.

[25] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (illustrated edition ed.). Yale University Press, New Haven.

[26] GLAAD. 2021. Social Media Safety Index 2021. https://www.glaad.org/blog/glaads-social-media-safety-index

[27] GLAAD. 2023. *Social Media Safety Index 2023.* Technical Report. GLAAD. 44 pages. https://assets.glaad.org/m/7adb1180448da194/original/Social-Media-Safety-Index-2023.pdf

[28] Kishonna L. Gray and Krysten Stein. 2021. "We 'said her name' and got zucked": Black Women Calling-out the Carceral Logics of Digital Platforms. *Gender & Society* 35, 4 (Aug. 2021), 538–545. https://doi.org/10.1177/08912432211029393

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

[30] Oliver L. Haimson, Justin Buss, Zu Weinger, Denny L. Starks, Dykee Gorrell, and Briar Sweetbriar Baron. 2020. Trans Time: Safety, Privacy, and Content Warnings on a Transgender-Specific Social Media Site. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–27. https://doi.org/10.1145/3415195

[31] Oliver L. Haimson, Avery Dame-Griff, Elias Capello, and Zahari Richter. 2019. Tumblr was a trans technology: the meaning, importance, history, and future of trans technologies. *Feminist Media Studies* 21, 3 (2019), 345–361. https://doi.org/10.1080/14680777.2019.1678505

[32] 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), 1–35. https://doi.org/10.1145/3479610

[33] 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

[34] Camille Harris, Amber Gayle Johnson, Sadie Palmer, Diyi Yang, and Amy Bruckman. 2023. "Honestly, I Think TikTok has a Vendetta Against Black Creators": Understanding Black Content Creator Experiences on TikTok. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–31. https://doi.org/10.1145/3610169

[35] @heycolanda. 2023. #FreeTheNipple. https://www.instagram.com/p/Cnnaz5xsNOr/?hl=en

[36] Anna Hoffmann and Anne Jonas. 2016. Recasting Justice for Internet and Online Industry Research Ethics. In *Internet Research Ethics for the Social Age: New Cases and Challenges*, Michael Zimmer and Katharina Kinder-Kurlanda (Eds.). Peter Lang Publishing, 27.

[37] Kai Jacobsen, Aaron Devor, and Edwin Hodge. 2022. Who Counts as Trans? A Critical Discourse Analysis of Trans Tumblr Posts. *Journal of Communication Inquiry* 46, 1 (Jan. 2022), 60–81. https://doi.org/10.1177/01968599211040835

[38] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–33. https://doi.org/10.1145/3359294

[39] Dogus Karabulut, Cagri Ozcinar, and Gholamreza Anbarjafari. 2023. Automatic content moderation on social media. *Multimedia Tools and Applications* 82, 3 (Jan. 2023), 4439–4463. https://doi.org/10.1007/s11042-022-11968-3

[40] Sonia Katyal and Jessica Jung. 2021. The Gender Panopticon: Artificial Intelligence, Gender, and Design Justice. *SSRN Electronic Journal* (2021). https://doi.org/10.2139/ssrn.3760098

[41] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–22. https://doi.org/10.1145/3274357

[42] Amber Madison. 2015. When Social-Media Companies Censor Sex Education. https://www.theatlantic.com/health/archive/2015/03/when-social-media-censors-sex-education/385576/

[43] Neil M. Malamuth. 1996. Sexually Explicit Media, Gender Differences, and Evolutionary Theory. *Journal of Communication* 46, 3 (Sept. 1996), 8–31. https://doi.org/10.1111/j.1460-2466.1996.tb01486.x

[44] Brandeis Marshall. 2021. Algorithmic misogynoir in content moderation practice. *Heinrich-Böll-Stiftung European Union* (June 2021). https://eu.boell.org/sites/default/files/2021-06/HBS-e-paper-Algorithmic-Misogynoir-in-Content-Moderation-Practice-200621_FINAL.pdf

[45] 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

[46] Alexander Monea. 2022. *The Digital Closet: How the Internet Became Straight*. The MIT Press. https://doi.org/10.7551/mitpress/12551.001.0001

[47] David Myles, Stefanie Duguay, and Lucia Flores Echaiz. 2023. Mapping the social implications of platform algorithms for LGBTQ+ communities. *Journal of Digital Social Research* 5, 4 (Sept. 2023), 1–30. https://doi.org/10.33621/jdsr.v5i4.162

[48] 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 *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 1246–1266. https://doi.org/10.1145/3593013.3594078

[49] Oversight Board. 2020. Breast cancer symptoms and nudity. https://www.oversightboard.com/decision/IG-7THR3SI1/

[50] Oversight Board. 2022. Securing ongoing funding for the Oversight Board. https://www.oversightboard.com/news/1111826643064185-securing-ongoing-funding-for-the-oversight-board/

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

[52] Oversight Board. 2023. Meet The Board. https://www.oversightboard.com/meet-the-board/

[53] Oversight Board. 2023. Oversight Board. https://www.oversightboard.com

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

[55] Oversight Board. 2023. Public Comment Appendix for 2022-009/10-IG-UA. https://oversightboard.com/attachment/574162714231347/

[56] Pinterest. 2023. Report nudity. https://help.pinterest.com/en/article/report-nudity

[57] 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 *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 1882–1895. https://doi.org/10.1145/3593013.3594134

[58] Audacia Ray. 2007. *Naked on the Internet: hookups, downloads, and cashing in on Internet sexploration*. Seal Press, Emeryville, CA. OCLC: ocm82367813.

[59] Real Facebook Oversight Board. 2023. Oversight Board Continues to Whitewash, Spin Its Flagging Impact In Annual Report; Summary Buries the Lede: 14.1% of Recommendations Actually Implemented by Meta. https://rfob.medium.com/oversight-board-continues-to-whitewash-spin-its-flagging-impact-in-annual-report-5709c93b5c15

[60] Sarah T. Roberts. 2019. *Behind the screen: content moderation in the shadows of social media*. Yale University Press, New Haven. OCLC: on1055263168.

[61] 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

[62] 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#:~:text=The%20actress%27%20account%20was%20quickly,women%20without%20her%20platform%20do%3F&text=When%20actress%20Trace%20Lysette%20made,profile%20banned%20with%20no%20explanation.

[63] Salty. 2019. Exclusive: An Investigation into Algorithmic Bias in Content Policing on Instagram. https://www.saltyworld.net/algorithmicbiasreport-2/

[64] 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

[65] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–33. https://doi.org/10.1145/3359246

[66] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 1–35. https://doi.org/10.1145/3392866

[67] 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

[68] Julia Serano. 2016. *Whipping girl: a transsexual woman on sexism and the scapegoating of femininity* (second edition ed.). Seal Press, Berkeley, CA. OCLC: ocn920728057.

[69] Shakira Smith, Oliver L Haimson, Claire Fitzsimmons, and Nikki Echarte Brown. 2021. Censorship of Marginalized Communities on Instagram, 2021. *Salty* (Sept. 2021). https://saltyworld.net/product/exclusive-report-censorship-of-marginalized-communities-on-instagram-2021-pdf-download/

[70] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability For Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–28. https://doi.org/10.1145/3476059

A APPENDIX

B DATA ANALYSIS CODEBOOK

Table 1: Data Analysis Codebook

Categories	Codes
Trans erasure in Meta's policy and moderation	Trans erasure, criticism of censorship, criticism of Meta's policies: not inclusive to all genders, lack of references to trans identities in Meta's policies, disproportionate moderation limits trans community building, criticism of Meta's nudity policy exceptions: does not encapsulate all aspects of gender confirmation processes
Vagueness of Meta's Adult Nudity and Sexual Activity policy	Criticism of Meta's nudity policy, clear nudity policies benefit human moderators, Meta's nudity policies do not account for grey areas, criticising vague wording of Meta's nudity policies, binary/bioessentialist policies make it unclear what trans users can post, criticising Meta's nudity policies as ineffective in practice
Inconsistent enforcement of Meta's Adult Nudity and Sexual Activity Policy and Sexual Solicitation Policy	Content inaccurately removed as adult/nudity, paid advertising rejected as adult or sexual content despite following guidelines, criticism of moderation: does not remove actual rule-breaking adult/nude content, criticism of algorithmic moderation: cannot accurately gender users
Trans visibility	Lack of trans visibility, importance of trans visibility, importance of visibility to trans youth, importance of internet/social media to trans youth, importance of Meta platforms for trans visibility, importance of queer visibility, importance of visibility to queer youth, comparing online trans erasure to LGBT book bans
Suggestions related to gender inclusivity and moderation	Allow users to post images including chests/nipples regardless of gender, policies should not be based on gender in general, either allow all or no images of nipples regardless of gender, use inclusive language about bodies, avoid using binary language to refer to bodies, stop embedding outdated views of gender in technologies, use non-gendered language for body parts in policy
Suggestions for clarifying policies	Nudity guidelines should be detailed, clearly define key policy terms, provide examples of content that do or don't comply with adult/nudity guidelines, provide clear nudity guidelines for moderators and users, clearly define difference between "appropriate" and "inappropriate" sexual content, remove vague/subjective words from policy, clarify moderation guidelines, stop censoring queer users, provide guidance for how to make intent re: sexual content clear to avoid incorrect removals, appeal process should be transparent and easily accessible, clearly explain how removed posts violate guidelines, provide easier/more effective appeal options for users