Recommend Me? Designing Fairness Metrics with Providers

Jessie J. Smith
Jessie.Smith-1@colorado.edu
University of Colorado, Boulder
USA

Aishwarya Satwani
aishwarya.satwani@colorado.edu
University of Colorado, Boulder
USA

Robin Burke
robin.burke@colorado.edu
University of Colorado, Boulder
USA

Casey Fiesler
casey.fielser@colorado.edu
University of Colorado, Boulder
USA

ABSTRACT

Fairness metrics have become a useful tool to measure how fair or unfair a machine learning system may be for its stakeholders. In the context of recommender systems, previous research has explored how various stakeholders experience algorithmic fairness or unfairness, but it is also important to capture these experiences in the design of fairness metrics. Therefore, we conducted four focus groups with providers (those whose items, content, or profiles are being recommended) of two different domains: content creators and dating app users. We explored how our participants experience unfairness on their associated platforms, and worked with them to co-design fairness goals, definitions, and metrics that might capture these experiences. This work represents an important step towards designing fairness metrics *with* the stakeholders who will be impacted by their operationalizations. We analyze the efficacy and challenges of enacting these metrics in practice and explore how future work might benefit from this methodology.

CCS CONCEPTS

• **Information systems**; • **Human-centered computing** → **User studies**;

KEYWORDS

Fairness, Recommendation, Focus Groups, Metrics

ACM Reference Format:
Jessie J. Smith, Aishwarya Satwani, Robin Burke, and Casey Fiesler. 2024. Recommend Me? Designing Fairness Metrics with Providers. 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/10.1145/3630106.3659044

1 INTRODUCTION

Whether or not a job ad gets applicants, a YouTube video earns ad revenue, a fundraiser meets its goal, or a dating profile gets any matches, depends largely on how much these things are seen. However, being "seen," or algorithmically *exposed* may be entirely at the mercy of recommender systems. Recommender systems leverage algorithms to recommend content, items, or information that matches users' imputed preferences. However, previous work has highlighted how these personalized systems might also lead to unintentional harm, such as degenerate feedback loops [37, 44], sexist stereotyping [33], or racial bias [5]. This realization has resulted in increased motivation to understand and improve "fairness" in recommendation systems.

Fairness is a complex, contested, contextual, and theoretical construct [35]. Fairness has been defined in many disciplines, with overlapping and sometimes conflicting categorizations both within and outside of machine learning (ML). Previous research has outlined dozens of fairness metrics that have been developed for recommendation and ranking applications and categorized them according to the underlying fairness goals and constraints that each metric is actually attempting to measure within the system [43, 49]. However, to the best of our knowledge, these metrics were not developed or designed with the people who *experience* fairness or unfairness on the platform. This presents an important gap in the discipline—although the goal of fairness measurement is to capture how fair a system is for its users, they are typically not included in the design of these metrics.

Therefore, in this work, we explore the lived experiences of fairness from the perspective of **providers** (people who provide items for recommendation), and co-design metrics that empirically capture these experiences of fairness. We provide a concrete method to assist ML practitioners with developing fairness definitions and metrics that are aligned with their users' needs, which previous work in the FAccT community has explicitly called for [15]. During four focus groups with thirteen total participants, we co-designed fairness metrics with providers from two domains of recommendation: content creators and dating app users. This cross domain analysis allowed us to compare and contrast fairness goals for providers who interact with different platforms and therefore likely have different algorithmic exposure needs. Specifically, our research questions were as follows:

- **RQ1:** What are the lived experiences and perceptions of unfairness for content creators and dating app users?
- **RQ2:** What fairness goals, definitions, and metrics do these providers want enacted in their recommender systems to alleviate these experiences of unfairness?
- **RQ3:** What are some of the opportunities and challenges of designing fairness metrics with providers from different recommendation domains and contexts?

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

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3659044

In the course of answering these questions, we learned that many experiences of unfairness across both domains of recommendation stemmed from algorithmic under-exposure, algorithmic over-exposure, and a lack of agency or understanding of how recommendation algorithms work. We also discovered several challenges and limitations of designing fairness metrics with providers, most notably that measuring fairness does not ensure fair outcomes, and that providers' individual preferences for fairness may conflict with one another or with an organization's goals.

The main contributions of this research are threefold. First, we identify several ways that content creators and dating app users experience unfairness on recommendation platforms. Second, we introduce a methodology for designing ML fairness metrics *with* users who are impacted by their operationalizations. This includes an analysis and mapping of participants' lived experiences of unfairness to their associated fairness goals, definitions, and metrics. Third, we analyze the consequences and tradeoffs that might occur if these metrics were enacted in a recommender system and discuss how future work might be able to better incorporate providers' perspectives into fairness metric design.

2 LITERATURE REVIEW & BACKGROUND

Fairness Operationalization is the process of scoping fairness goals, defining fairness based on those goals, selecting an appropriate fairness metric based on this definition, and incorporating it into the objectives of an ML system [35, 49, 53]. When operationalizing fairness for ML systems, most literature is focused on mathematical [8] or statistical [13] definitions of fairness. This growing interest to research fairness in machine learning through statistical approaches has led to a proliferation of fairness definitions and metrics that machine learning practitioners can use to empirically measure how fair or unfair their systems are [6, 40]. One application of fairness in machine learning is in the context of recommender systems (and information access systems more generally [23]).

2.1 Measuring Fairness in Recommendation

Recommender systems are used as a way to filter information to people. The key, user-side stakeholders of a recommender system are the **consumers** (end-users who receive the recommendations) and the **providers** (people who provide items that will be recommended) [9]. Identifying fairness concerns for each of these stakeholders, and then translating those concerns into empirically measurable constructs could yield a variety of different approaches—some of which might even come into conflict with each other [1, 50].

Fairness in recommendation considers questions such as: *Are we fairly representing the information space? Are we giving all item providers an opportunity to be seen or engaged with? And are we recommending items to users in a way that fairly distributes information or resources to those users?* [23]. Operationalizing these concerns into fairness metrics is challenging in part because fairness goals are amorphous and heavily context-dependent [41]. In addition, it is notably difficult to ensure construct validity while designing fairness metrics. Construct validity is concerned with answering the question: *are we actually measuring the thing that we are trying to measure?* [42]. Although recent work has begun to explore how to assess construct validity for fairness measurements in ML [35], there has been less focus on assessing the validity of fairness measurements from the perspective of users. Despite the underlying goal of measuring fairness—to empirically capture how fairly or unfairly users are being treated by a system—users themselves are rarely involved in the process of developing fairness metrics for ML systems.

2.2 Lived Experiences of Fairness

When fairness metrics are designed without input from users, they more often align with fairness definitions from moral philosophy (e.g., utilitarianism [50]) or from policy (e.g., the Civil Rights Act of 1964 [4]). However, users' lived experiences of fairness might differ from these *theoretical* fairness definitions. Instead, users' perceptions of whether or not a platform is treating them fairly are often based on their algorithmic experiences or their folk theories. **Algorithmic experience** is a version of user experience that is specifically focused on the users' interactions with and perceptions of algorithms [38]; while **Folk Theories** are informal theories that users develop about how systems function, which influence their perceptions and interactions with those systems [16, 18]. Understanding users' algorithmic experiences and folk theories can help practitioners develop better algorithm and platform designs to align the algorithmic functionality with users' expectations [16, 38, 46]. In general, asking users about their experiences and perspectives of fairness with recommendation platforms is a very promising research direction to help create more fair platforms, and one that has garnered research interest in recent years.

2.3 Providers' Perspectives of Fairness

Eliciting stakeholders' perspectives about what they consider to be fair algorithmic treatment is an important step toward designing fairness metrics. Previous work has explored both *consumer's* perspectives on recommendation fairness (e.g., [3, 47, 52, 54]) and *provider* perspectives (e.g., [21, 22, 25, 26, 45]). Although both consumer and provider perspectives are useful, Jannach and Bauer [36] describe that, while the original intention of recommender systems was to make life easier for consumers, they ought to also prioritize and create value for providers. In this work, we consider the perspectives of providers from two domains: dating app users and content creators.

2.3.1 Perspectives From Dating App Users. In the context of dating apps, several studies have explored the kinds of harm that can occur to LGBTQ+ users. This research has uncovered that when profile recommendations do not match the user's specified matching preferences, this can lead to hateful messaging, identity erasure, and feelings of unsafety [7, 27, 28]. Another fairness concern for dating apps is that they might prioritize recommending users based on their perceived level of "attractiveness," even at the cost of more compatible matches [11]. This algorithmic design choice could capture and even perpetuate social inequities, such as sexual racism [34].

2.3.2 Perspectives From Content Creators. One algorithmic fairness concern for content creators is that algorithms' objectives do not always align with creators' objectives. For example, recommendation algorithms sometimes conflict with the creative process [12],

or reward behavior that runs counter to creativity [48] (e.g., by rewarding frequent posting, or forcing creators to recycle old videos to recreate virality). Another fairness concern for content creators is related to recommendation-based content moderation. Gillespie [29] explains how recommendations can be used as a means to moderate content, reduce harmful content exposure, or shadowban (algorithmically blocking or suppressing content without transparency). Previous research has shown that political conservatives, transgender people, and Black people have their content and accounts banned most often [31], and are more likely to have their content labeled "unsafe" [39]. This kind of unfairness has left content creators feeling confused, frustrated, and powerless [32, 55].

2.4 Designing Fairness with Users

In response to all of these experiences and perceptions of unfairness, researchers have begun to turn towards participatory and collaborative methods to better align recommendation algorithms with users' needs [14, 19, 52]. Jannach and Bauer [36] describe how conducting research with users is important to validate the success of a recommender system, rather than optimizing the system for metrics without understanding how those operationalizations might impact real people. In addition, Stray et al. [53] describe how designing metrics with system stakeholders can help make systems more robust, and better aligned with human values. In this work, we explore this promising method of fairness design by assessing providers' lived experiences of fairness and co-designing fairness metrics to capture those experiences.

3 METHODS

This study consisted of four virtual focus groups with a total of thirteen participants, which is in line with standards for user studies in HCI [10]. Focus groups were the most appropriate method to adopt for this work because of our goals to uncover tensions between providers and to encourage brainstorming that extended beyond participants' individual experiences through group discussions. In addition, focus groups encourage research participants to develop ideas collectively, and sharing experiences among participants can prompt people to elaborate on their stories and themes, which can help researchers interpret those experiences [51]. The first two focus groups included participants who are content creators on platforms like TikTok, Instagram, or YouTube. The final two focus groups included participants who use dating apps. This research was approved by our university's institutional review board.

3.1 Recruitment

We recruited our participants via open calls to participate on Twitter, LinkedIn, Instagram, TikTok, and Slack channels with personal networks. Interested participants were sent to an online form to select which type of provider they were. All participants were compensated $30 USD for their participation. Since the goal of this study was to elicit the perspectives and preferences of recommendation *providers*, we asked participants to concentrate on this role during the focus groups. For example, even though content creators become consumers when they scroll others' posts or dating app users become consumers when they swipe on others' profiles—both of these groups are also providers (users whose content/profile is exposed to others). To ensure we elicited preferences of the provider, we asked participants to engage in activities while remembering their vested interest in getting their content or profile exposed to other users. We chose to focus on providers because their perspective is often missing from recommender system design and evaluation and also because, especially in the case of platforms where providers can be paid, there are higher stakes for providers.

All participants and their collected attributes are included in Table 1 and Table 2. We included gender identity and race/ethnicity in the tables as self-described by participants, to preserve their preferred terminology. We note the lack of diversity in both gender identity and racial/ethnic identity of our recruited participants as a limitation of this research study, and our results should be interpreted with this in mind. The majority female, majority White demographic of our participants likely omits useful dimensions of unfairness experienced by other demographic groups, but still provides a helpful starting point for this seminal work. Each participant was assigned an alias based on their recommendation domain. For example, content creators were assigned the alias PC#, whereas dating app users were assigned the alias PD#. (e.g., "PC1" or "PD1"). We use these aliases as reference throughout the remainder of this paper. Through our pilots, we found that limiting focus groups to 3-4 participants was ideal to allow for active participation.

3.2 Focus Group Design

The focus group design was the same for all four focus groups, regardless of the recommendation domain in question. All participants were guided to an online collaboration document on Google Docs. Each activity involved independent brainstorming in the shared document, as well as a group discussion afterward. All focus groups were conducted virtually via Zoom and were recorded on the same platform. Audio recordings were then transcribed using Microsoft Word's audio transcription software. The focus groups included 4 different activities for participants:

(1) **Introduction Activity.** Participants brainstormed about their experiences with the recommender systems that they are a provider for, as well as their perception of the algorithms on these platforms. Then all participants introduced themselves and shared with one another.
(2) **Unfairness Activity.** Participants brainstormed about their experiences of unfairness and fairness as providers. Participants were not provided with a definition of "fairness" or "unfairness," and were asked to recall their experiences based on their personal perceptions of what fair treatment was for them. After brainstorming, participants volunteered to discuss their experiences, while the research team took notes to come up with a "fairness concerns list" that captured these experiences.
(3) **Fairness Goals and Definitions Activity.** Participants collectively selected a recommendation platform in their domain that everyone was familiar with (e.g., Hinge for dating app users, or Instagram for content creators). They brainstormed a list of fairness goals related to this platform, with specific attention towards the fairness concerns list that was

Table 1: Collected demographics from content creator participants. FG# refers to the associated focus group for each participant.

Alias	FG#	Type of Content	Platform(s) They Post Content On	Cumulative Follower #	Age Range	Gender Identity	Race / Ethnicity
PC1	1	Art & sewing	Instagram, YouTube	Between 500-1000	20-30	Female	White
PC2	1	Short-form and long-form animation	LinkedIn, Vimeo, Tumblr, Instagram, YouTube	Less than 500	20-30	Nonbinary	White
PC3	1	Historical education	YouTube, Instagram, TikTok, Facebook, Threads, Patreon	More than 10,000	30-40	Female	White/ Jewish
PC4	2	Movement practitioner videos	Instagram, TikTok, Facebook	More than 10,000	20-30	Female	White/ Hispanic
PC5	2	AI-facilitated digital artist	TikTok, Instagram	Less than 500	50-60	Male	White/ Non-Hispanic
PC6	2	Artist/Illustrator	Instagram, Facebook	Between 4,000 and 10,000	40-50	Female	White

Alias	FG#	Dating App(s) They Use	Paid for Premium?	Age Range	Gender Identity	Race / Ethnicity
PD1	3	Tinder, Bumble, OkCupid, Hinge, Taimi, Her, Lex, Grindr, Feeld, Facebook Dating	No	20-30	Female	White/ Latine
PD2	3	Hinge, Tinder	Yes	20-30	Male	South Asian (Indian)
PD3	3	Hinge, Tinder, Bumble	Yes	20-30	Nonbinary/ Agender	White
PD4	4	Tinder, Bumble	No	20-30	Female	White
PD5	4	Hinge, Tinder, Feeld	No	20-30	Female	White
PD6	4	Hinge, Bumble, Tinder, The League, Coffee Meets Bagel	Yes	20-30	Female	White
PD7	4	Bumble, Hinge, Feeld, Tinder	No	20-30	Cisgender Woman	White/ Caucasian

Table 2: Collected demographics from dating app participants. FG# refers to the associated focus group for each participant.

created during the previous activity. Participants also developed a condensed fairness definition that attempted to capture these fairness goals. Together, the group discussed their fairness goals and definitions, as well as their challenges and experiences with the activity.

(4) **Fairness Metrics Activity.** Using the same platform as decided in the previous activity, all participants brainstormed at least one "Fairness Metric" that allowed them to measure whether or not one of their fairness goals were being achieved by the recommendation system. They were asked to include details about (1) which data would need to be collected to measure fairness in this way; (2) which user populations (demographics) might need to be compared against one another; (3) how they would know if "fairness" had been achieved; and (4) how they would know if the platform was still not "fair" enough. Together, the group discussed their experiences with trying to design fairness metrics, including identifying any tradeoffs and challenges with their specific metric(s), and how they would like this process to be done in practice.

3.3 Data Analysis

Using the transcribed audio recordings of the focus groups, the first author conducted a version of thematic analysis via inductive open coding, by tagging individual sections and quotes with labels that were associated with that particular observation. Written responses from participants during brainstorming sessions were also coded. When a new response or quote fit into a previously defined code, it was included in that category. After creating a codebook with related quotes and observations, the first two authors and the final author discussed the observations and codes and came to a consensus about emerging themes. These high-level themes are discussed throughout the remainder of this paper.

4 RESULTS

During our focus groups, we first discovered that most participants were uncertain about how recommendation systems worked. This lack of understanding also influenced participants' abilities to recount their experiences of algorithmic fairness. For example, PD5 said *"I think it's kind of hard to say whether [I'm] being treated fair or unfairly just because I don't really have a good understanding of how the algorithm works overall."* PD4 also shared that they felt *"uncertain about the level of fairness,"* that they experience, *"just because of*

the opacity of the like algorithm itself." Although uncertainty about the algorithm's functionality made it difficult at times to understand if the recommender system was being fair, participants still had many experiences of unfairness to share and discuss with one another.

In the following sections, we describe these experiences and how they informed the participants' designs of their fairness goals, definitions, and metrics. Though there were commonalities between the two domains, overall there were many contextual differences, so we categorized these results by the participant type (Content Creators or Dating App Users) and noted when any similarities were observed between both groups.

4.1 Lived Experiences of Unfairness

When asking participants about their experiences of fairness and unfairness on their associated platforms, several participants struggled with interpreting the word "fair," such as PC1, who expressed that it's *"hard to say what is unfair, just what **feels** unfair"* (PC1). As noted in Section 3.2, we did not provide a definition of fairness for participants, nor did we explicitly elicit one from participants. Although this decision prevented us from ensuring that responses aligned with a mutually agreed understanding of what "fairness" is, this absence of a definition also allowed us to explore what our participants considered to be "unfair" based solely on their personal experiences and preferences.

4.1.1 ***Content Creators.*** For content creators, one experience of unfairness that emerged was inequality of content virality or exposure. PC1 shared that they felt the algorithm prioritized creators with a large following, regardless of their quality of content: *"I think you know the sort of narrative that... if you [make] good content that people want to see, you will go viral... however I see many of times that big creators do the same as small creators and go viral for it"* (PC1).

PC5 expressed discouragement that their profile might never be fairly exposed by the algorithm, given their small follower count: *"when I look at the profiles of other people who are doing the same thing, they frequently have 10's of thousands of followers. I don't feel like I'm ever going to get there when I only get one or two new followers per day"* (PC5).

Another experience of unfairness recounted by our participants was related to algorithmic content suppression and shadowbanning (algorithmically blocking/suppressing provider's content without their knowledge) [29]. *"My observation of Instagram is that it's suppressing everybody all the time. When they switched to reels it was clear that they had no interest in pushing image content at all"* (PC3). PC4 and PC6 both shared experiences where their content was removed for being *"too sexual,"* even though it was them *"dancing fully clothed"* (PC4). These participants further discussed how their content had been banned (PC6) or algorithmically suppressed (PC4) without explanation. PC6 noted one instance where a post was removed *"about [not] blaming victims of sexual assault and telling them that they need to be aware that they could get assaulted on the subway,"* and that they were not sure why this content was banned.

When these participants were asked *why* these experiences felt unfair, they described that it led to feelings of frustration (P3) and upset (P6), or caused them to lose opportunities when their content was incorrectly suppressed without any warning or rectification (P4). In summary, for the content creators, unfairness was mostly experienced when their content did not receive enough exposure on the platform, or when their content was suppressed by the recommendation algorithm—without any understanding as to why this was happening, nor the agency to prevent it.

4.1.2 ***Dating App Users.*** For the dating app participants, one shared experience of unfairness was related to unwanted profile exposure that did not align with their stated preferences. For example, PD1 said that *"I want to see [women] in my dating apps, and somehow men keep seeing my profile and keep liking my profile in a way where I genuinely don't understand, like how that happens"* (PD1). This participant said that this mismatch between preferences and exposure felt unfair because it was wasting both their time and their prospective matches' time, and possibly leading to a loss of opportunity. *"If there's people who are like a categorical rejection for me, like, I will categorically reject men on dating apps, then it's like wasting my time. It's wasting their time... it kind of feels like something has been sort of like taken away, or like you've lost an opportunity"* (PD1).

Similarly, PD3 shared that even though they turned off preferences for men on their Hinge app, they still received likes from men and wondered *"how the hell did you even see me? Because I'm, like, turning [that preference] off. Like I don't want to talk to men"* (PD3). This participant said the inability to expose their content based on their explicitly stated preferences felt unfair to them because it seemed like the algorithm was making assumptions about their identity and sexual preferences based on their appearance, which felt binarist.

> *It feels unfair to me. Like I present extremely feminine, and I know that... But like, I feel like it's taking the way I look, the way I present for my picture... and then completely filtering out an entire like group of people by saying 'well like even if someone is nonbinary, if they look a certain way, we're going to put them in the same category as [men or women]'. It just feels like it's binarist without saying it is, even if it gives you that third gender option* (PD3).

PD2 added an additional consequence of this fairness concern; when their profile had been exposed to people outside of their stated preferences, this led to hateful speech in their direct messages. *"I had one time this very racist person send... something like very racist... So that was weird, like if my preference is not [this person] why are you sending me likes from these people?* (PD2). PD5 similarly described that when their profile was exposed to people who did not align with their preferences, they would *"get ignored or kind of talked down to,"* and this led them to feel like the platform was *"not really a safe space"* (PD5).

In summary, all of these experiences showed that when dating profiles are algorithmically exposed to an incorrect audience, this can lead to feelings of discomfort, loss of opportunity, feelings of discrimination or exclusion, and concerns for safety on the platform. Interestingly, it appeared that many fairness concerns for content creators were about under-exposure, while many fairness concerns for dating app users were about over-exposure—implying

that fairness goals, definitions, and metrics designed to capture these experiences might need to measure different effects.

4.2 Fairness Goals and Definitions

After discussing the participants' experiences of unfairness on their associated platforms, we asked them to complete a series of activities to develop fairness goals and definitions that might improve their experiences of fairness on a given platform. Examples of some of the fairness goals and definitions that participants developed are shown in Table 3. We discovered two categories of fairness goals and definitions that were shared between dating app users and content creators: (1) exposure equality; and (2) transparency.

4.2.1 ***Content Creators.***

- ***Exposure Equality.*** Several participants designed fairness goals and definitions to improve equality of content exposure. For example, PC4's fairness definition was to provide *"equal opportunity for all [creators]"*, while one of PC3's fairness goals was to improve *"exposure consistency"* (PC4). PC1 similarly defined fairness as not *"demonetiz[ing] shops or business accounts, [don't] reward people with high views or daily content [and] stop suppressing creators that aren't white men, women who show skin, or non-cis creators."* PC1 and PC2 (who had discussed that it felt unfair when their followers weren't being recommended their content) designed fairness goals that prioritized exposing content to followers first. PC1 felt that *"people who have already indicated interest in a creator's content—followers, liked— [should] be shown that creator's content as priority"* (PC1).
- ***Transparency.*** PC4 thought that fairness could be improved through greater transparency around *"policies, especially content revoking reasoning,"* while PC3 described that fairness could be improved through transparency about *"what factors into a successful post"* and *"algorithmic changes."*

4.2.2 ***Dating App Users.*** Although transparency and exposure equality were also fairness goals for dating app users, the nuances of these goals differed in this domain.

- ***Exposure Equality.*** PD1, PD3, PD4, PD5, and PD7 all mentioned that every user should have an equal opportunity to be shown to prospective dates, regardless of their identity or dating preferences. PD3 additionally included that there should be equality in the allocation of benefits: *"every user should gain benefit from selection processes, showing and being shown to preferences."* PD2 also included a fairness goal to not promote *"any sort of racism, sexism, ableism, homophobia, transphobia, [casteism], and religious discrimination"* (PD2). This kind of exposure equality was also described as a *"utopian"* fairness goal because it would seek to *"expos[e] the user to all types of individuals to either match the user's existing perspective or to broaden it"* (PD5).
  In contrast, several participants described exposure equality as aligning someone's profile exposure with their explicitly shared dating preferences, such as PD1 who said *"I think like actually following through on... the settings that you put in and then actually like honoring those is like a very basic first step [towards fairness]."* PD7 also took this a step further and described a fairness goal where the algorithm should *only* take into account explicitly stated preferences, rather than implicitly observed ones: *"follow explicit user preferences, but not implicit ones (e.g. making assumptions based on who the user has swiped on in the past)"* (PD7).
- ***Transparency.*** Several participants detailed how improving transparency could increase the *feeling* of fairness through added agency on the platform, such as PD3 who said, *"transparency would go a really, really long way to making an app feel more fair to me."* This participant further shared that, *"getting a little bit more of a glimpse into how things work would go a long way toward making it seem more fair.* ***Even if it isn't.****"* This implied that even the *appearance* of transparency might make recommendation algorithms feel more fair based on the users' experience, regardless of whether or not the algorithm is *theoretically* fair. PD2, PD4, and PD5 also designed fairness goals related to improving transparency and agency surrounding why their profile is shown to others, and why certain profiles are shown to them.

In summary, although both content creators and dating app users developed fairness goals and definitions that promoted increased transparency and exposure equality, the nuances about which types of exposure would feel fair differed between these two recommendation domains.

4.3 Fairness Metrics

In the final activity of the focus groups, we asked participants to develop their own fairness metrics, based on their fairness goals and definitions. Each participant brainstormed what the goal of their metric was, the kinds of data they might need to measure this in practice, and the fairness "threshold" their metric would need to meet to consider the system fair enough for deployment. Here we describe several of these metrics for both recommendation contexts.

4.3.1 ***Content Creators.***

- ***Content Quality and Exposure Equality Metrics.*** Both PC3 and PC5 developed a metric to measure the quality of someone's content. These participants thought content exposure should only rely on the content *quality*, not on identity attributes related to the creator. *"The algorithm shouldn't discriminate among factors that are not related to the content. For example, race and gender identity should not be a factor for content about art"* (PC5). To operationalize this metric, PC3 thought it would be necessary to collect demographic data (e.g., race, culture, language, gender identity, and personal aesthetics) to evaluate exposure differences between these attributes. This participant explained that if they were to use this metric, they would know that the platform is unfair if *"we still see small creators who make quality content getting very few views, seeing little to no growth in followers and engagement"* (PC3). PC4 developed a similar fairness metric that sought to measure equality of content exposure, based on demographic groups of content creators. They described that *"if all accounts [from different demographic groups] have a similar exposure percentage average (within 5% of each other), fairness is achieved"* (PC4).

Alias	Fairness Goal(s)	Fairness Definition	Fairness Metric
PC1	• *"Do not censor female/non male/non white/non cis bodies"*	*"Stop suppressing creators that aren't white men, women who show skin, or non-cis creators."*	Exposure Equality Metric
PC4	• *"Exposure consistency"* • *"Transparency around policies, especially content revoking reasoning"*	*"Providing equal opportunity for all users."*	Exposure Equality Metric
PD1	• *"Profiles have equal reach"* • *"Profiles have accurate audience"* • *"Profiles of marginalized identities not excluded from/within normative profiles"*	*"We will ensure that all profiles have equal visibility within the spaces/ groups/ populations the user desires to be seen with."*	Categorical Rejection Rate Metric
PD5	• *"Show a wide variety of individuals to mimic a real life setting"* • *"Show my profile to a wide variety of individuals"*	*"Mimic the outside world in its utopian state, exposing the user to all types of individuals to either match the user's existing perspective or to broaden it."*	Hotness Metric
PD7	• *"Users should be shown profiles (and have their profile shown) to a variety of people"* • *"Follow explicit user preferences, but not implicit ones (e.g. making assumptions based off of who the user has swiped on in the past)"* • *"Allow users to adjust their preferences and profile and adjust algorithm behavior accordingly"*	*"Follow stated user preferences while also allowing for diversity of shown preferences, and constantly adjust so as not to pigeon-hole any user."*	Diversity Metric

Table 3: Examples of fairness goals and their associated definitions and metrics designed by participants.

- ***Popularity Bias Metric.*** Another metric developed by both PC2 and PC1 evaluated fairness for content creators with a small following. This metric is aligned with measuring the concept of Popularity Bias, the phenomenon in recommender systems where popular items receive most of the algorithmic exposure, while less popular items remain systematically under-exposed [2]. PC2 thought that this metric could measure what percentage of someone's social media feed is from profiles with small follower counts, and could be used to optimize exposure for those kinds of accounts. PC1's metric sought to improve popularity bias by measuring if certain content receives more or less exposure when posted from small accounts versus large accounts.

4.3.2 *Dating App Users*.

- ***Transparency Metric.*** PD3 and PD4 developed a version of a transparency fairness metric. PD3's transparency metric was related to a questionnaire that users could fill out on a dating app. Their idea was that this metric could measure how compatible profiles are to one another and that fairness would be achieved if this information was shared with users. PD4's transparency metric instead focused on how much of the algorithms' functionality was being explained and shown to users—if every profile sorting and matching mechanism was being shown to users, this would be considered fair.
- ***Categorical Rejection Rate Metric.*** PD1 developed a fairness metric that would *"identify how many categorically incompatible profiles are being shown to a user,"* with a specific focus on improving compatible match performance for marginalized (e.g., LGBTQ+) users. PD2 developed a similar metric that sought to measure how much someone's profile exposure aligned with their explicitly stated preferences, specifically with respect to gender and sexuality. This participant determined that fairness would be "achieved" for this metric, *"if a provider's profile is presented to more than at least 50-60% of their intended target user"* (PD2).
- ***Diversity and Hotness Metrics.*** Finally, both PD5 and PD7 developed metrics to try to improve dating discrimination being perpetuated through dating apps. PD5 developed a *"hotness metric"* that sought *"to expose individuals to all different types of people with varying appearance [and] to not prioritize profiles that have received more likes/swipes or who appear more stereotypically attractive"* (PD5). PD7 developed a similar metric that sought to *"measure if the profiles being shown to a user are diverse across multiple features visible in their profile... within the user's stated explicit preferences"* (PD7). For this fairness metric, PD7 described that they thought the platform would be deemed fair enough, *"if every user's recommended profiles score >=50% diversity."*

4.3.3 *Who Should Design Fairness Metrics?*

We also asked participants who they would ideally like to be involved in this process of designing fairness metrics for the recommender systems that they interact with. In all four focus groups, participants noted that the practitioners involved in this work should come from diverse backgrounds, which included a diversity of culture, geographical background, gender identity, sexual preferences, age, race, relationship status, and disciplinary background. PC1, PC3, PC6, PD2, and PD5 all specifically requested diversity for the programmers who might implement fairness metrics into the system. PD4 requested that fairness metrics should be designed by people who have used

the apps before. PD1 requested that critical researchers, or *"people whose research focuses on marginalized identities"* be included. Finally, many participants (PC2, PC6, PC1, PC4, PD2, and PD6) also expressed that they would like fairness metrics to be designed with the *users* of these algorithmic systems.

4.4 Challenges with Measuring Fairness

During the focus groups, participants described various tradeoffs that might occur if their metrics were operationalized, as well as challenges that they faced when attempting to design fairness metrics generally.

4.4.1 ***Content Creators.*** PC2 shared that their goal to prioritize exposing content to followers might help content creators, but could also unintentionally lead to filter bubbles. *"It does kind of create a bit of a bubble... if you're seeing [just] the things that your followers are seeing"* (PC2). PC1 thought that their fairness metric (to measure what percentage of a creator's followers are exposed to their content) might only work well for content creators with a small number of followers, but not work as well for creators with large follower counts.

Another major challenge arose during a lively discussion between PC5 and PC6 when both participants realized that their fairness needs could not be met simultaneously on Instagram. PC6 described how, as an artist, they felt it would be fairer to ban AI-generated art from social media: *"I see the work of people I know who has been stolen into [AI-generated art] and these people are saying they're getting less and less views and that makes me very angry because obviously that's theft. And Facebook and Instagram are not making a ban on [AI-generated art], and that's really bad because I think they should be protecting [artists] … and we cannot opt out of [generated content], which is really unfair"* (PC6).

In contrast, PC5, who creates AI-generated art, noted that banning or algorithmically suppressing that content would feel unfair to them. In this example, both PC5 and PC6 had different experiences of fairness and unfairness. Banning or suppressing AI-generated art might improve fairness for PC6 at the cost of fairness for PC5. Alternatively, exposing and recommending AI-generated art might improve fairness for PC5 at the cost of fairness for PC6.

4.4.2 ***Dating App Users.*** The main tradeoff that emerged for dating app users was a tension between increasing diversity of dating app recommendations while also preserving safety and explicitly stated preferences. PD7 described how increasing diversity of profile recommendations might also increase hate speech: *"there have been trans people who use [dating] apps and get matched with people who are transphobic and then like they get hate crimed on this app."* PD5 similarly expressed concern about increasing the diversity of recommendations on dating apps. They shared that this would be *"making the assumption that all people are kind and respectful and accepting, and it's just not the truth"* (PD5).

This led to a discussion about the responsibility of dating apps in general, where participants began to question if it is a dating app's responsibility to stop or hinder dating discrimination through fairness operationalization, even though it exists offline. PD5 thought it was the responsibility of dating apps to attempt to portray the world in its most utopian state. *"Let's portray a world that has like so much fairness, so much love. No racism … like explosive diversity. I think … it's the company's responsibility"* (PD5). PD6 similarly thought that it should be the responsibility of dating apps to not perpetuate any kind of discrimination or harm, but that certain apps could cater to specific preferences that already exist in the real world: *"A lot of Muslim girls I know use Minder … people can have their biases and have a whole app for that. You know they can design for that in a non-harmful way for people who want to date certain types of people"* (PD6).

PD4 took this a step further and noted that *"there's a part of me that's like this whole [dating] process is inherently unfair, and that's sort of part of it. [Dating apps are] trying to make it too fair, [and] maybe [it's] not actually what people want"* (PD4). Ultimately, the participants agreed that this tradeoff would be inevitable when operationalizing fairness for dating apps, and did not know how to remedy this challenge. *"A lot of times one of these things that could benefit one group can harm another group and it's hard to balance those"* (PD7).

Another set of challenges related to the efficacy of measuring fairness in practice. One example mentioned by PD1 and PD2 was the difficulty of deciding what their fairness "threshold" should be (how to determine which cutoff for their metric should be considered fair or unfair). *"What are the kind of arbitrary numbers that we're choosing to denote success or denote failure?"* (PD1). This participant also expressed concern that using fairness metrics in general might lead practitioners to believe they are measuring and optimizing for fairness when their measurements might not be accurately capturing users' lived experiences of fairness.

> *My concern... is that developers are going to take [metrics] as like the final step and they're going to kind of stop caring as long as they can keep this one certain numerical metric satisfied... they're not going to care to put resources towards more subjective, qualitative experiences of unfairness* (PD1).

PD3 added to this concern and felt like metrics might lead platforms to further exclude and marginalize certain users, just for the sake of good PR. *"If your populations are [small] enough... if you can make it as inhospitable as possible to the demographic that you're trying to have … fairness for … Then you're going to hit that [fairness threshold] every time, because there's no one there"* (PD3). All of these challenges highlight an important reflection about whether fairness ought to be measured at all. PD6 felt that fairness might be empirically immeasurable, even through proxies like fairness metrics: *"I thought [designing fairness metrics] was very hard to do. Thinking about how you can make these things empirical or like proving them empirically... it seems like an impossible task, honestly"* (PD6). We further explore these challenges of measuring fairness in the following section.

5 DISCUSSION

Here we unpack some of the opportunities and challenges of designing fairness metrics with providers from different recommendation domains and contexts. We also discuss how future research might best adopt this methodology.

5.1 Opportunities and Challenges with Fairness Metric Design

Goals Versus Outcomes. As PD1 and PD3 described during their focus group, the *goal* to measure and optimize a recommender system for fairness might not necessarily guarantee the *outcome* of fairness for providers. This concern has been previously introduced through Goodhart's law—the notion that when a metric becomes a target, it ceases to be a good measure [30]. In line with this challenge, PD6 also mentioned their concern that fairness measurement might become a PR goal rather than a real effort to improve users' lived experiences of fairness. This also introduces a challenge around fairness measurement generally—should recommender systems be attempting to operationalize a certain theoretical kind of fairness? Or should they be optimizing for users' lived experiences and perceptions of fairness? Previous work has shown that recommender systems can still be perceived as unfair by users, even when the generated recommendations are theoretically fair [24]. Thus, in future work, ML practitioners will likely need to discern what *kind* of fairness their system is attempting to measure, and what the end goal of that measurement should be.

Inherent Tradeoffs. Another major challenge observed during focus groups was the inherent tradeoffs that might exist when operationalizing certain fairness metrics over others. One example of this arose among the dating app users when our participants expressed two competing desires: the desire to increase the diversity and exposure of their dating profiles (to decrease dating discrimination); and the desire to limit profile exposure to align with explicitly declared preferences (to decrease discomfort and unsafety on the platform). Both of these fairness goals were in direct conflict with one another, which could make it impossible to operationalize both goals at once. Another example of this kind of tradeoff emerged during the discussion between PC5 and PC6, where the decision to ban or suppress AI-generated content might make the platform feel more fair for some providers while making the platform feel less fair for others.

Domain Specificity. Throughout these focus groups, we learned that some fairness goals, definitions, and metrics were shared between content creators and dating app users. Both groups of participants were interested in improving transparency and exposure equality on their respective platforms. However, the nuances of how these goals might be measured or enacted differed between domains. For example, exposure equality for content creators PC1, PC4, and PC5 required that the algorithm prioritize recommendations based on the *quality* of someone's content. Previous work has shown that musicians also feel that not all music content is equally deserving of exposure and that algorithmic exposure should be based in part on quality [21]. We note that exposing items based on their quality is an open challenge in recommendation research; although measures such as *expected exposure* from information retrieval attempt to capture this notion [20], there are still many challenges with attempting to accurately measure item quality in practice. In contrast to content creators, exposure equality for dating app users PD1, PD3, PD4, PD5, and PD7 required that the algorithm give equal (and accurate) exposure to everyone, regardless of the "quality" of their profile. This difference implies that although different recommendation domains might have similar fairness goals, the operationalizations of these goals might need to be domain-specific.

Generalizing Between Domains. For recommender systems, increased item exposure is sometimes thought of as a fairness guarantee [2]. However, by exploring providers' experiences of unfairness, we have learned that increased item exposure could, at times, actually lead to increased *unfairness* instead. Although several content creators in our study wanted more exposure of their content overall–previous work has shown that certain marginalized groups of creators do not want their content exposed to the "wrong" audience, for safety reasons. For example, DeVito [17] discovered that trans creators wanted the algorithm to stop exposing their content to transphobic users because this could lead to hate speech. Similarly, in our study, PD1, PD2, PD3, and PD5 all shared that they did not want their dating profile exposed to the "wrong" audience, because it could lead to feelings of discrimination, exclusion, discomfort, unsafety, or loss of opportunity. These fairness goals, if shared between domains, could potentially be operationalized in similar ways.

5.2 Limitations & Future Work

Here we outline two limitations of our approach to designing fairness metrics, and how they might impact future work that adopts this methodology. First, when eliciting fairness concerns from users, we recommend balancing this evidence against what else is deemed "fair" for a given system. Users might be able to provide useful knowledge about their personal preferences towards fairness measurement, but these may not be compatible with the constraints or underlying goals of the system. In other words, certain systemic fairness concerns might actually run counter to what individual stakeholders want. For example, previous work has highlighted that dating apps may have a duty to disobey users' explicitly stated preferences at times, because those preferences might reflect and even amplify societal discrimination [34]. However, in our study, we learned that some dating app participants felt it was unfair if the algorithm ignored their explicitly stated preferences. When an organization's fairness goals conflict with users' individual experiences of fairness, our method of eliciting fairness metric design from users' preferences might be less appropriate.

Second, throughout this research, we noticed that there may be desirable normative properties of recommender systems that are not necessarily "fairness" properties. For example, many of our participants pointed towards transparency as a component of fairness, which is not a normative claim about a distribution of benefits. Transparency, instead, serves as a check on a system so that stakeholders can verify it is acting in the way they expect. However, we can still learn from users' personal representations of fairness, regardless of whether they align with theoretical notions of fairness, because they serve as a good reminder that fairness is not the only desirable property to strive for when making socially beneficial recommender systems.

6 CONCLUSION

In this study, we conducted four focus groups with thirteen providers of recommender systems to understand how these stakeholders

might design their own fairness metrics based on their lived experiences of algorithmic unfairness. We learned that dating app users and content creators have experienced a breadth of unfairness on their respective platforms, including algorithmic under-exposure, algorithmic over-exposure, and a lack of agency. Participants were able to develop their own fairness goals, definitions, and metrics to try to capture these experiences of unfairness, and learned that there may not be a measurement method that can benefit all users at once. We hope this work acts as a helpful case-study for practitioners who would like to design fairness metrics and interventions with the users who will be impacted by them, and encourage future work to further explore this kind of design methodology.

ACKNOWLEDGMENTS

We thank Samantha Dalal, Zoe Fisher, Joshua Paup, Ellen Simpson and members of TRSL and IRL for their help with piloting, data shares, and research feedback. This research was supported in part by the National Science Foundation under award IIS-2107577 and by Google Research.

RESEARCH ETHICS AND SOCIAL IMPACT

Ethical Considerations

As human subjects research, this study was approved by our institutional IRB and thus complied with current best research ethics practices, which included the following:

- All authors of this paper have completed a CITI certification, which outlines best practices for human subjects research.
- Voluntary and informed consent was obtained from every participant.
- Participants were each compensated $30 USD for participating in the 90-minute focus group.
- Participants were told at the beginning of every focus group that participant identities were to remain confidential, and should not be shared outside of the context of the research study.
- Participants were allowed to leave the focus group at any time they deemed necessary.
- All participant data was stored on a password-protected server and was only available to the authors of this paper. Transcripts from the focus groups were anonymized with participant aliases immediately after recording and before coding was conducted. We also ensured that no identifiable information was included in the reporting of these results.

In addition to these practices, the first author of this paper (who was the lead facilitator for the focus groups) began every session with a script that warned participants about the potential emotional challenges that might occur while describing lived experiences of unfairness. The facilitator encouraged participants to support one another and to communicate to other participants with respect, to ensure that the focus groups would remain a safe space for sharing and collaboration.

Positionality

The authors of this paper represent multiple genders and races, as well as marginalized identities that were raised in this study's data. A subset of the authors also have personal experiences as dating app users and content creators.

Adverse and Unintended Impact

Although we do not anticipate adverse impacts of this work, one potential unintended impact is worth noting. As we described in Section 5.2, our method to incorporate users' perspectives into the design of fairness metrics could conflict with methods to improve systemic fairness concerns in recommendation systems. We note this potential impact in the paper, and encourage future research to critically examine if designing fairness metrics with users is appropriate for their context.

REFERENCES

[1] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. 2020. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction* 30 (2020), 127–158.

[2] Himan Abdollahpouri and Masoud Mansoury. 2020. Multi-sided exposure bias in recommendation. *arXiv preprint arXiv:2006.15772* (2020).

[3] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2020. The connection between popularity bias, calibration, and fairness in recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 726–731.

[4] Civil Rights Act. 1964. Civil rights act of 1964. *Title VII, Equal Employment Opportunities* (1964).

[5] Julia Angwin and Terry Jr. Parris. 2016. Facebook lets advertisers exclude users by race. https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2018. Fairness and Machine Learning. fairmlbook. org, 2019.

[7] Rena Bivens and Anna Shah Hoque. 2018. Programming sex, gender, and sexuality: Infrastructural failures in the "feminist" dating app Bumble. *Canadian Journal of Communication* 43, 3 (2018), 441–459.

[8] Meredith Broussard. 2023. *More than a Glitch: Confronting Race, Gender, and Ability Bias in Tech*. MIT Press.

[9] Robin Burke. 2017. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093* (2017).

[10] Kelly Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 981–992.

[11] Musa Eren Celdir, Soo-Haeng Cho, and Elina H Hwang. 2023. Popularity bias in online dating platforms: Theory and empirical evidence. *Manufacturing & Service Operations Management* (2023).

[12] Yoonseo Choi, Eun Jeong Kang, Min Kyung Lee, and Juho Kim. 2023. Creator-friendly Algorithms: Behaviors, Challenges, and Design Opportunities in Algorithmic Platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–22.

[13] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (2020), 82–89.

[14] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–23.

[15] Wesley Hanwen Deng, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio. 2023. Investigating Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 705–716.

[16] Michael Ann DeVito. 2021. Adaptive folk theorization as a path to algorithmic literacy on changing platforms. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–38.

[17] 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 (2022), 1–31.

[18] Michael A DeVito, Darren Gergle, and Jeremy Birnholtz. 2017. " Algorithms ruin everything" # RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 3163–3174.

[19] Michael Ann DeVito, Ashley Marie Walker, and Julia R Fernandez. 2021. Values (mis) alignment: Exploring tensions between platform and LGBTQ+ community design values. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–27.

[20] Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. 2020. Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 275–284.

[21] Karlijn Dinnissen and Christine Bauer. 2023. Amplifying Artists' Voices: Item Provider Perspectives on Influence and Fairness of Music Streaming Platforms. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. 238–249.

[22] Karlijn Dinnissen, Isabella Saccardi, Marloes Vredenborg, and Christine Bauer. 2023. Looking at the FAccTs: Exploring Music Industry Professionals' Perspectives on Music Streaming Services and Recommendations. In *Proceedings of the 2nd International Conference of the ACM Greek SIGCHI Chapter*. 1–5.

[23] Michael D Ekstrand, Anubrata Das, Robin Burke, Fernando Diaz, et al. 2022. Fairness in information access systems. *Foundations and Trends® in Information Retrieval* 16, 1-2 (2022), 1–177.

[24] Mehdi Elahi, Himan Abdollahpouri, Masoud Mansoury, and Helma Torkamaan. 2021. Beyond algorithmic fairness in recommender systems. In *Adjunct proceedings of the 29th ACM conference on user modeling, adaptation and personalization*. 41–46.

[25] Andres Ferraro, Xavier Serra, and Christine Bauer. 2021. Break the loop: Gender imbalance in music recommenders. In *Proceedings of the 2021 conference on human information interaction and retrieval*. 249–254.

[26] Andres Ferraro, Xavier Serra, and Christine Bauer. 2021. What is fair? Exploring the artists' perspective on the fairness of music streaming platforms. In *IFIP Conference on Human-Computer Interaction*. Springer, 562–584.

[27] Lindsay Ferris and Stefanie Duguay. 2020. Tinder's lesbian digital imaginary: Investigating (im) permeable boundaries of sexual identity on a popular dating app. *New Media & Society* 22, 3 (2020), 489–506.

[28] Avery Garritano. 2021. Inside a binary interface: the construction of gender and identity in mainstream dating apps. *Occam's Razor* (2021).

[29] Tarleton Gillespie. 2022. Do not recommend? Reduction as a form of content moderation. *Social Media+ Society* 8, 3 (2022), 20563051221117552.

[30] Charles Goodhart. 1975. Problems of monetary management: the UK experience in papers in monetary economics. *Monetary Economics* 1 (1975).

[31] 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 (2021), 1–35.

[32] 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 (2023), 1–31.

[33] HRW. 2018. *" Only Men Need Apply": Gender Discrimination in Job Advertisements in China*. Human Rights Watch.

[34] Jevan A Hutson, Jessie G Taft, Solon Barocas, and Karen Levy. 2018. Debiasing desire: Addressing bias & discrimination on intimate platforms. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–18.

[35] Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 375–385.

[36] Dietmar Jannach and Christine Bauer. 2020. Escaping the mcnamara fallacy: towards more impactful recommender systems research. *Ai Magazine* 41, 4 (2020), 79–95.

[37] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. 2019. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 383–390.

[38] Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. 2021. Algorithmic folk theories and identity: How TikTok users co-produce Knowledge of identity and engage in algorithmic resistance. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–44.

[39] Sara Kingsley, Proteeti Sinha, Clara Wang, Motahhare Eslami, and Jason I Hong. 2022. " Give Everybody [..] a Little Bit More Equity": Content Creator Perspectives and Responses to the Algorithmic Demonetization of Content Associated with Disadvantaged Groups. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–37.

[40] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proc. conf. fairness accountability transp., new york, usa*, Vol. 1170. 3.

[41] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency*. 39–48.

[42] Paul C Price, Rajiv S Jhangiani, and I-Chant A Chiang. 2015. Reliability and validity of measurement. *Research methods in psychology* (2015).

[43] Amifa Raj and Michael D Ekstrand. 2020. Comparing fair ranking metrics. *arXiv preprint arXiv:2009.01311* (2020).

[44] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 131–141.

[45] Vishal Sharma, Kirsten E Bray, Neha Kumar, and Rebecca E Grinter. 2022. Romancing the Algorithm: Navigating Constantly, Frequently, and Silently Changing Algorithms for Digital Work. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–29.

[46] Ellen Simpson, Andrew Hamann, and Bryan Semaan. 2022. How to Tame" Your" Algorithm: LGBTQ+ Users' Domestication of TikTok. *Proceedings of the ACM on Human-computer Interaction* 6, GROUP (2022), 1–27.

[47] Ellen Simpson and Bryan Semaan. 2021. For You, or For" You"? Everyday LGBTQ+ Encounters with TikTok. *Proceedings of the ACM on human-computer interaction* 4, CSCW3 (2021), 1–34.

[48] Ellen Simpson and Bryan Semaan. 2023. Rethinking Creative Labor: A Sociotechnical Examination of Creativity & Creative Work on TikTok. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[49] Jessie J Smith, Lex Beattie, and Henriette Cramer. 2023. Scoping Fairness Objectives and Identifying Fairness Metrics for Recommender Systems: The Practitioners' Perspective. In *Proceedings of the ACM Web Conference 2023*. 3648–3659.

[50] Jessie J Smith, Anas Buhayh, Anushka Kathait, Pradeep Ragothaman, Nicholas Mattei, Robin Burke, and Amy Voida. 2023. The Many Faces of Fairness: Exploring the Institutional Logics of Multistakeholder Microlending Recommendation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1652–1663.

[51] Janet Smithson. 2008. Focus groups. *The Sage handbook of social research methods* (2008), 357–370.

[52] Nasim Sonboli, Jessie J Smith, Florencia Cabral Berenfus, Robin Burke, and Casey Fiesler. 2021. Fairness and transparency in recommendation: The users' perspective. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 274–279.

[53] Jonathan Stray, Alon Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Chloe Bakalar, Lex Beattie, Michael Ekstrand, Claire Leibowicz, et al. 2022. Building human values into recommender systems: An interdisciplinary synthesis. *ACM Transactions on Recommender Systems* (2022).

[54] Yen Nee Wong, Rhia Jones, Ranjana Das, and Philip Jackson. 2023. Conditional trust: Citizens' council on data-driven media personalisation and public expectations of transparency and accountability. *Big Data & Society* 10, 2 (2023), 20539517231184892.

[55] Jing Zeng and D Bondy Valdovinos Kaye. 2022. From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet* 14, 1 (2022), 79–95.