Group Fairness for Content Creators: the Role of Human and Algorithmic Biases under Popularity-based Recommendations

Ionescu, Stefania ; Hannak, Aniko ; Pagan, Nicolo

DOI: https://doi.org/10.1145/3604915.3608841

Group Fairness for Content Creators: the Role of Human and Algorithmic Biases under Popularity-based Recommendations

Stefania Ionescu
ionescu@ifi.uzh.ch
University of Zürich
Zürich, Switzerland

Anikó Hannák
hannak@ifi.uzh.ch
University of Zürich
Zürich, Switzerland

Nicolò Pagan
nicolo.pagan@uzh.ch
University of Zürich
Zürich, Switzerland

ABSTRACT

The Creator Economy faces concerning levels of unfairness. Content creators (CCs) publicly accuse platforms of purposefully reducing the visibility of their content based on protected attributes, while platforms place the blame on viewer biases. Meanwhile, prior work warns about the "rich-get-richer" effect perpetuated by existing popularity biases in recommender systems: Any initial advantage in visibility will likely be exacerbated over time. What remains unclear is how the biases based on protected attributes from platforms and viewers interact and contribute to the observed inequality in the context of popularity-biased recommender systems. The difficulty of the question lies in the complexity and opacity of the system. To overcome this challenge, we design a simple agent-based model (ABM) that unifies the platform systems which allocate the visibility of CCs (e.g., recommender systems, moderation) into a single popularity-based function, which we call the *visibility allocation system* (VAS). Through simulations, we find that although viewer homophilic biases do alone create inequalities, small levels of additional biases in VAS are more harmful. From the perspective of interventions, our results suggest that (a) attempts to reduce attribute-biases in moderation and recommendations should precede those reducing viewers' homophilic tendencies, (b) decreasing the popularity-biases in VAS decreases but not eliminates inequalities, (c) boosting the visibility of protected CCs to overcome viewers' homophily with respect to one fairness metric is unlikely to produce fair outcomes with respect to all metrics, and (d) the process is also unfair for viewers and this unfairness could be overcome through the same interventions. More generally, this work demonstrates the potential of using ABMs to better understand the causes and effects of biases and interventions within complex sociotechnical systems.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing theory, concepts and paradigms**; **Social network analysis**.

RecSys '23, September 18–22, 2023, Singapore, Singapore

ACM ISBN 979-8-4007-0241-9/23/09.
https://doi.org/10.1145/3604915.3608841

KEYWORDS

algorithmic fairness, agent-based modeling, network formation, popularity bias

ACM Reference Format:
Stefania Ionescu, Anikó Hannák, and Nicolò Pagan. 2023. Group Fairness for Content Creators: the Role of Human and Algorithmic Biases under Popularity-based Recommendations. In *Seventeenth ACM Conference on Recommender Systems (RecSys '23), September 18–22, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3604915.3608841

1 INTRODUCTION

The steep growth of the Creator Economy [4, 10, 13, 19] also brought increasing concerns regarding embedded inequalities. As millions of content creators (CCs) now earn livable wages by posting content on platforms such as Instagram, YouTube, and TikTok [13, 18], they naturally expect fair opportunities and remuneration in return. Despite this, the market shows severe inequalities based on protected attributes. For instance, figures show a 30% gender earning gap and 35% racial earning gap [27, 38, 44], each paired with imbalances in the follower counts [24, 48].

Several lawsuits [2, 3] and public declarations [23, 50] indicate that CCs believe platforms to be at least partially at fault. The main accused problem is the artificial decrease in the visibility of content from protected CCs either as a result of recommender systems (e.g., by amplifying existing imbalances in the data [11, 37]) or moderation (which aims to ensure the compliance with the policies of platforms but, e.g., may allegedly contain discriminatory guidelines [23] or result in unfair culling of content [3]). As a response, platforms claim these *attribute biases* are not hard-coded into the algorithms and are most likely the effect of the algorithms picking up on the viewers' biases. All those concerns come in the context of the well-documented issue of and efforts to reduce *popularity-bias* (i.e., the tendency to recommend more popular CCs more, thus reinforcing their popularity) in recommender systems (RSs) [1, 11, 34, 51, 54]. The feedback loop generated by the popularity bias likely exacerbates any existing attribute biases of viewers and platforms.

In this paper, we tackle one question that remains unclear: *Which of the above-mentioned factors are the most harmful, and what interventions would be the most effective?* The difficulty of the question lies in the complexity of the system, the differences between platforms, and the lack of access to the algorithms, which makes it hard to address it via surveys or online experiments. Thus, we use agent-based modeling (ABM), a methodology that proved effective for exploring causal effects of various complex systems with many intercommunicating individuals [20]. Examples include understanding the long-term effects of ML-enhanced decision-making systems [12], inequalities in ride-sharing platforms [6], the effect

of interventions that reduce biases on online dating platforms [25], and the effects of RSs in diverse domains [33, 40, 53]. Tied to our application domain, Pagan et al. [43] and Ionescu et al. [26] used a simple ABM, where viewers follow CCs solely based on the quality of the content, to better understand how popularity-biased RSs contribute to the lack of predictability of success empirically observed in cultural markets [14, 45][1].

In this paper, we investigate the aforementioned question, with a focus on *group unfairness* (i.e., inequalities between protected and unprotected groups [5, 15]). First, we alter the ABM of Pagan et al. [43] by adding a protected attribute that allows for accounting for the homophily of viewers [29]. Then, we formalize three group fairness metrics for both CCs and viewers. Finally, we simulate the resulting system[2] and capture the previously defined metrics in order to understand (a) what is the individual and joint role of viewers' and platform's biases in the inequality of the system for CCs, (b) which CCs are affected the most, (c) what are the most effective interventions, (d) what are potential side-effects or caution points for these interventions, and (e) how do biases and interventions affect viewers.

2 MODEL

To address our research question, we need a simple, interpretable model that isolates recommendations from the decision-making process of viewers. The closest one is the quality-based model [43], which we extend for our scope by incorporating protected attributes (e.g. race, gender). The synthetic platform consists of n content creators (CCs) and m *seekers* (i.e., users who focus exclusively on consuming the content). The process unfolds by iterations corresponding to one unit of time (e.g., one day). Each iteration consists of two phases: (a) the platform suggests one CC to each seeker, and (b) each seeker decides whether or not to follow the recommended CC. From a network perspective, the platform starts with an empty network. At each iteration, edges from seekers to CCs are added, thus always maintaining a bipartite structure of the network[3].

2.1 The decision-making process of seekers

In each iteration, seekers decide whether or not to follow the suggested CC based on their valuation of this CC. Namely, they follow the suggested CC only if they assess it as having more value than any of the CCs followed thus far. The valuation is based on the attributes of both the seeker making the decision and the considered CC.

Attributes of CCs. Each CC c has two attributes: quality (q_c) and type (t_c). Quality is an attribute on which seekers would generally agree (e.g., video resolution, clarity of sound) and would all prefer to maximize. Type, however, is an attribute whose assessment depends on personal taste (e.g., serious versus funny videos, gender), and seekers prefer to match on. This distinction is anchored in prior work which differentiates between ordinal (or competing) and nominal (or matching) attributes [7, 25, 32, 47]. For simplicity, we include only one quality and one (binary) type attribute. For our goal, type represents a protected attribute (e.g., race), but the model can be extended beyond this assumption.

Attributes of seekers. Each seeker s has a type t_s and a level of bias b_s. The type is chosen from the same set as the type of CCs and dictates what sort of creators the seeker prefers (e.g., race-wise). The level of homophilic bias dictates how much importance the seeker puts on finding a CC with a matching type.

Valuation. The attributes of both the seeker and the considered CC determine how much a seeker values the content of that CC. As done in prior agent-based models (e.g., in school choice [16] or online dating [25]), we assume the final score is a linear combination between quality and type-match, which is dictated by the bias level of the seeker who makes the assessment. Formally, the value a seeker s gives to CC c is given by the valuation function $v_s(c) = (1-b_s)\cdot q_c + b_s\cdot(\mathbb{1}_{t_c=t_s} - \mathbb{1}_{t_c\neq t_s})$. This valuation induces a ranking, $\succ_s$, of each seeker over the existing CCs.

Example. Assume that CC c has quality $q_c = 3$ and type $t_c = 1$, and another CC c' has quality $q_{c'} = 1$ and type $t_{c'} = 0$. Next, let a seeker s of type $t_s = 0$ give a level of importance of $b_s = 2/3$ to having a matching type. Then that seeker would value the first CC by $v_s(c) = (1-2/3)\cdot 3 + 2/3\cdot(0-1) = 1/3$ and they would value the second CC by $v_s(c') = (1-2/3)\cdot 1 + 2/3\cdot(1-0) = 1$. Therefore, the seeker prefers c' over c, i.e., $c' \succ_s c$ despite $q_{c'} < q_c$.

2.2 Platform Suggestions - the Visibility Allocation System (VAS)

Recommendation function. In the first phase of an iteration, the platform suggests one CC to each seeker. This can be represented as a mapping which we call the *recommendation function*. We formally denote it by $R : S \to C$ where S is the set of seekers and C is the set of CCs.

Types of recommendations. The suggestions are made based on the current state of the network. We refer to the network at iteration t by a^t, which is an m by n matrix with $a_{s,c}$ being 1 when seeker s follows CC c and 0 otherwise. The final recommendations are, in practice, the result of several processes, including the moderation process and the recommender system (RS). For simplicity, we refer to the resulting system producing the recommendation function as the *visibility allocation system* (VAS). This paper uses the same three systems defined in [26]. They represent common network formation processes that vary the popularity bias (from none to extreme):

- UR The uniform random VAS recommends each CC with equal probability, i.e., $\mathbb{P}(R_{UR}(s) = c) = \frac{1}{|C|}, \forall c \in C, s \in S$.
- PA The preferential attachment VAS gives the more popular CCs higher visibility: $\mathbb{P}(R_{PA}(s) = c) = \frac{1+a_{\cdot,c}}{|C|+a_{\cdot,\cdot}}, \forall c \in C, s \in S$, where dots represent summation (i.e., $a_{\cdot,c}$ is the total number of followers of CC c and $a_{\cdot,\cdot}$ is the sum of the number of followers of all CCs).
- EPA The extreme preferential attachment VAS picks a uniform random recommendation among the most followed CCs:

[1] To ensure this short paper remains brief, we integrated the related work relevant for positioning and motivating the paper within the introduction and the one relevant for informing the model design and calibration within the remaining sections.

[2] The code is publicly available on our git repository: github.com/StefaniaI/ABM-GFforSMI.

[3] Even though usually any user can share content and be followed, we make the simplifying assumption that only a small fraction of users do. This was also assumed in the prior model [26, 43] and observed in practice [42]. We also take the number of seekers to be much larger than the number of CCs.

for all seekers $s \in S$, $\mathbb{P}(R_{EPA}(s) = c) = \frac{1}{|\arg\max_{c \in C} a_{\cdot,c}|}$ if $c \in \arg\max_{c \in C} a_{\cdot,c}$, 0 otherwise.

While prior work shows that PA is a realistic type of network-formation process [41], UR and EPA are extreme examples of VAS that are unlikely to be encountered in practice as they are. However, UR and EPA proved helpful in understanding what happens when more or less exploratory versions of PA are implemented [26].

Biased recommendations. Throughout the past few years, CCs accused platforms of having biased VAS which reduced the visibility of protected CCs e.g. by the unfair culling of videos during moderation [2, 3]. We implement this type of disadvantage by considering a biased alternative for each of the three VAS above. To reflect the view of CCs, we say a system biased at an l%-level will reduce the visibility of protected CCs by making each follower count l% less. Formally, let $g_l : C \to \mathbb{R}$ be a function which is 1 for unprotected CCs, and $1 - l$ for protected ones[4]. Then any biased system generates recommendation functions simply by multiplying visibility weights by this g_l function:

- Biased UR: $\mathbb{P}(R_{Biased\_UR_l}(s) = c) = \frac{g_l(c)}{\sum_{c \in C} g_l(c)}$.
- Biased PA: $\mathbb{P}(R_{Biased\_PA_l}(s) = c) = \frac{(1+a_{\cdot,c}) \cdot g_l(c)}{\sum_{c \in C}(1+a_{\cdot,c}) \cdot g_l(c)}$.
- Biased EPA: $\mathbb{P}(R_{Biased\_EPA_l}(s) = c) = \frac{g_l(c)}{\sum_{c \in \arg\max a_{\cdot,c} \cdot g_l(c)} g_l(c)}$ if $c \in \arg\max_{c \in C} a_{\cdot,c} \cdot g_l(c)$ and 0 otherwise.

3 FAIRNESS METRICS

To address our research question, we first build on existing literature on algorithmic fairness [9, 52] to define group unfairness, which we measure as the gap in either normalized cumulative success (for CCs) or satisfaction (for seekers) between members of the protected and unprotected groups. More precisely, given a metric μ for CCs, the level of unfairness at time t with respect to μ is $U^t_\mu = \frac{1}{|C_{\neg p}|} \cdot \mu_{C_{\neg p}}(a^t) - \frac{1}{|C_p|} \cdot \mu_{C_p}(a^t)$, where $C_p \subseteq C$ is the set of protected CCs and $C_{\neg p} \subseteq C$ is the set of unprotected CCs. Analogous for seekers. As such, positive (negative) values correspond to outcomes that disadvantage protected (unprotected) CCs or seekers. We acknowledge the diversity of metrics for success and satisfaction which capture different key aspects of outcomes. Thus, we do not limit our analysis to one metric alone. We include both measures taken at a precise time step and at *convergence* (i.e., when, for any possible recommendations, no seeker would follow a new CC)[5]. We denote by a^∞ the network at convergence.

For CCs in $X \subseteq C$, we measure their success by the following three metrics:

- The expected number of followers at convergence, i.e. $\mu_X(a^\infty) = \sum_{c \in X} \mathbb{E}[a^\infty_{\cdot,c}]$;
- The expected visibility at different timesteps, t, i.e. the expected number of seekers the CCs in X were recommended to, $\mu_X(a^t) = \sum_{c \in X} \mathbb{E}[|\{s : R^t(s) = c\}|]$;
- The chance of obtaining an individual fair (IF - see below) outcome, i.e. $\mu_X(a^\infty) = \mathbb{E}[|\{c \in X : a^\infty \text{ is IF for } c\}|]$.

While the first two metrics are general and self-explanatory, the latter is more deeply rooted in the issues within the Creator Economy. Similarly to the job market, an individually fair (IF) system would give better outcomes to more deserving individuals. This is why prior work defines an outcome as being IF for the i-th best CC if that CC is in the top i according to the number of followers [26]. In a simplified setting where quality is the only attribute, the authors proved that while with respect to the expected number of followers (i.e., ex-ante) the outcomes are IF for all CCs, it is likely that the actual realized outcome will not be fair for most CCs. The likelihood depends on the level of popularity bias within the recommendations. Based on these observations from prior work, the third metric looks at how group unfairness is distributed with respect to the fraction of CCs who receive an individually fair outcome and would thus believe they have been treated fairly with respect to the quality of their content.

Although our main focus is the fairness of CCs, seekers are also impacted by interventions, so we also track fairness with respect to them. Since the most common metrics for consumers refer to their level of dissatisfaction, we reverse-score the satisfaction of seekers. More precisely, we track the dissatisfaction of a subset of seekers $X \subseteq S$ by:

- The expected search time at timestep t [28, 49], i.e. the number of timesteps until finding the most preferred CC, $\mu_X(a^t) = -\sum_{s \in X} \mathbb{E}[|\{t : a^t_{s,\text{top}(\succ_s)} = 0\}|]$, where $\text{top}(\succ_s)$ is the best CC according to the preferences of seeker s;
- The expected quality of recommendations at timestep t, i.e. the position of the recommended CC within the preference of the seeker, $\mu_X(a^t) = -\sum_{s \in X} \mathbb{E}[\text{pos}_{\succ_s}(R^t(s))]$, where $\text{pos}_{\succ_s}(c)$ is the position of CC c in the ranking $\succ_s$ of the seeker s;
- The chance of seekers to be recommended a CC of a unmatching type, i.e., $\mu_X(a^t) = -\mathbb{E}[|s \in X : t_s \neq t_{R^t(s)}|]$.

4 THE VIRTUAL EXPERIMENT

Interventions. The experiments are designed to compare the effects of (i) reducing the level of bias (homophily) of seekers and (ii) modifying the visibility allocation system (VAS). First, we start with a biased population of seekers with homophilic preferences and a biased VAS and simulate the effects of reducing each. Since overcoming user biases is a long-term process [31], next we investigate whether systems that increase the visibility of protected CCs (i.e., have a negative level of bias, l) could overcome the inequalities resulting from seekers' biases. If so, we ask how much compensation is needed in order to observe fair outcomes and whether there are side effects.

Parameters. Although minimalistic, our model has a few fixed and variable parameters which define the users and the environment (see Table 1). Since our model has randomness, we run each parameter configuration with 500 distinct random seeds. We chose this number to allow for reasonable run times while giving reliable estimates of the outcome variables of interest. Therefore, we measure the results of the same experimental setup for different random seeds. We report their mean and say two measures (e.g., the success

[4]We interpret negative values of l as VAS which advantage the protected CCs by boosting their follower count with $|l|$%, i.e. $g_l : C \to \mathbb{R}$ is $1 + |l|$ for protected CCs. We will use them to model interventions targeting the VAS.

[5]Convergence is eventually reached for any VAS as the set of recommendable CCs weakly decreases with time and seekers will not follow any new CC after they were recommended the best one from this set (see Ionescu et al. [26] for a formal proof in the unidimensional setting).

Fixed parameters	Values
Number of CCs (n)	50
Number of seekers (m)	1000
Protected attribute (t_c, t_s)	$\in \{0, 1\}$
% protected (seekers and CCs)	25%
Varied parameters	**Values**
Level of seekers' bias (b_s)	$\in [0, 1]$
Fraction of biased seekers (f)	$\in [0, 1]$
Level of visibility allocation bias (l)	$\in [-1, 1]$
Visibility allocation system (VAS)	(Biased) UR, PA, EPA

Table 1: Tabular description of model parameters and their possible values during simulations.

of CCs with respect to the number of followers with and without a given intervention) differ *significantly* if their means are at least the sum of their standard deviations apart. We also performed a sensitivity analysis that goes beyond the one reported in the upcoming Results section. This showed that parameter variations (e.g., altering the level of bias of users) have a limited impact (e.g., only slightly raises unfairness) and do not change the takeaways of the paper. Additional sensitivity analysis is in our GitHub repository.

Choices around parameters. We emphasize that our model is a schematic representation of the system and thus resembles models focusing on understanding the causes of inequalities [46] more than those focusing on predictions [39]. Despite this, we aimed to ground our parameter choices within reality. First, we center our results around PA, as prior work showed it is a realistic network-formation process both in general [41] and specifically to CC-centered platforms [43]. Second, none of the VAS alternatives produce personalized recommendations; since the only matching attribute is protected, doing so would mean purposefully transferring the homophily of seekers into CC-discriminatory recommendation functions. This goes precisely against current efforts of debiasing RSs [11]. Third, the fraction of protected users varies with respect to both the studied platform and the considered attribute [19], so we chose an average value of 25%. We do not vary it within this report as our sensitivity analysis showed only predictable changes in the magnitude of the unfairness: Decreasing (increasing) the size of the minority value exacerbates (diminishes) the inequalities, especially for the top-quality CCs. Finally, prior work suggests that the user biases depend on exposure and could thus be lowered through interventions [17, 35]. Therefore, we model the biases of seekers by two parameters, namely their level of bias (b_s) and the fraction of seekers who exhibit such a bias (which we denote by f).

5 RESULTS

Recommendation and moderation biases have a higher impact than seeker biases. The first round of simulations investigates the effects of seeker and visibility allocation system (VAS) biases on the degree of unfairness for CCs. The results, which are depicted in Figure 1, carry three important takeaways. First, they suggest that biases in VAS pose more concerns about the level of unfairness than seekers' biases. As expected, when neither the VAS nor any of the seekers are biased, the level of unfairness is small. From that point onward, increasing either the level of VAS bias l or the share of homophilic seekers f produces unfairness. However, even when all seekers have homophilic biases ($f = 1$) the level of unfairness is lower than the one corresponding to an $l = 25\%$ VAS bias. Second, and perhaps surprisingly, when VAS are biased against protected CCs, reducing the fraction f of seekers with homophilic biases could exacerbate the unfairness for CCs. The simplicity of the model allowed for an interpretation of the cause: When fewer seekers are homophilic, there are in particular fewer protected seekers who will not follow unprotected CCs no matter how much extra visibility they get. Those first two observations suggest that bias-reducing interventions in recommender systems and moderation are of higher priority than those at the seekers' level. Finally, out of the three VAS, PA results in the highest levels of unfairness. While increasing the popularity bias (EPA) seems to reduce group unfairness, we know from prior work that doing so creates extreme levels of individual unfairness for the top-quality CCs [25]. Combined, these two findings support the increased push to look beyond accuracy and aim for exploration, diversity, and a decrease in popularity bias in RSs [22, 30, 36].

Seekers' biases impact top- and bottom-quality CCs differently. Next, we look at whether, depending on their quality, CCs are affected differently by biases and interventions. Figure 2a shows that higher rates f of homophilic biased seekers significantly decrease the average number of followers of the protected top-quality CCs, while maintaining a similar number of followers for the remaining CCs. VAS's biases l, however, lead to unfair outcomes for protected CCs disregarding where they situate quality-wise (see Figure 2b). The final plot of Figure 2 shows that reducing the fraction f of homophilic biased seekers when VAS are also biased ($l = 0.5$) only decreases the average number of followers of protected CCs, especially among the top-quality ones. This increases the level of unfairness for protected CCs. On the other hand, reducing the bias l in allocation systems decreases the unfairness at the top and eliminates the one for the remaining CCs. These results once again suggest that interventions targeting recommendations and moderation should come before those targeting the homophily of viewers. Moreover, it shows that eliminating biases in recommender systems and moderation could even lead to group fair results for two-thirds of CCs.

Boosting the visibility of protected CCs could overcome seeker biases; however, outcomes are not fair with respect to all metrics. Figure 1 showed that the homophily of seekers leads to inequality even in the absence of explicit biases in VAS. Moreover, it suggests that interventions that increase the visibility of protected CCs (i.e., a negative level of bias) could be used successfully to reduce unfairness. The level at which the platform needs to intervene depends largely on different factors, such as the level of popularity bias in recommendations. What remains unknown is whether or not such an intervention can lead to fair outcomes with respect to all three metrics of CC fairness. The first plot in Figure 3 shows that an intervention at a -25% level leads to a fair outcome with respect to the average number of followers. However, the next plot shows that while unfairness exists with respect to the allocated visibility too, a -25% level intervention has limited effectiveness. More precisely, while the intervention helps reduce long-term unfairness, it significantly disadvantages CCs from the unprotected group in

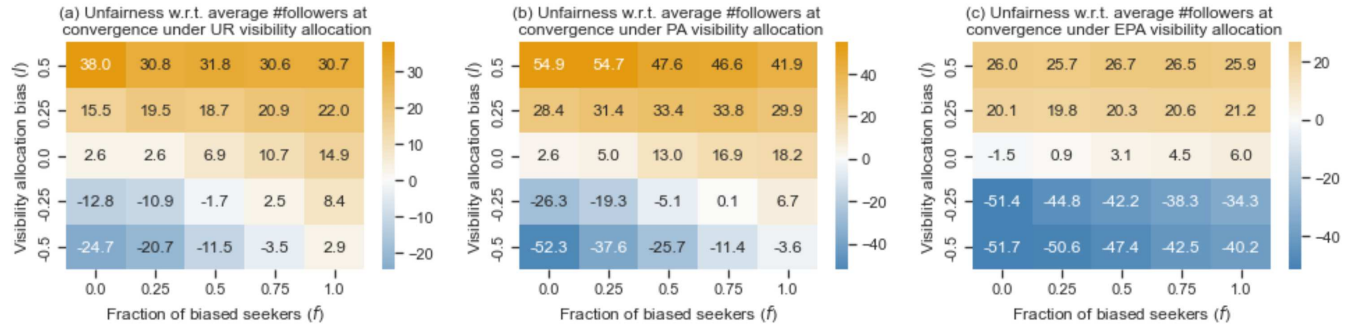

Figure 1: Each plot shows the unfairness with respect to the average number of followers at convergence for different levels of visibility allocation biases (l) and different shares of seekers with homophilic biases (f). Positive values (orange) correspond to outcomes unfair for protected CCs, while negative values (blue) correspond to outcomes unfair for unprotected CCs. We show the results for the three different VAS in increasing order of popularity biases. Throughout, biased seekers have medium biases (i.e., $b_s = 0.5$).

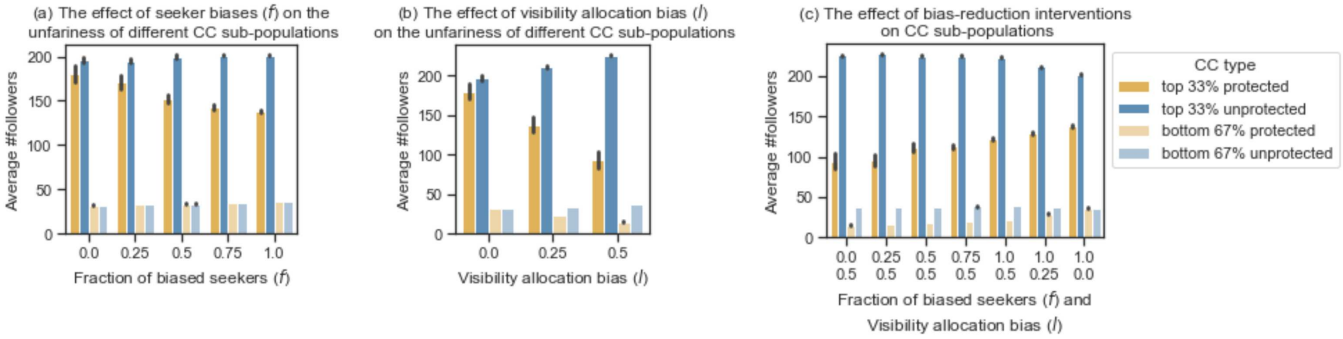

Figure 2: The bar plots show the impact of seeker and visibility allocation biases in producing unfairness (measured as the average number of followers) in the top-quality third and lowest-quality two-thirds of the CC population. From left to right, plots show (a) the effect of seeker population biases (f) when recommendations are not biased ($l = 0$), (b) the effect of recommendation biases (l) when none of the seekers are biased ($f = 0$), and (c) starting from a scenario where all seekers are biased (fifth point on the x-axis), the effects of reducing the fraction of biased seekers (progressing to the left) and of reducing the level of visibility allocation bias (progressing to the right). All plots are for PA recommendations and whenever seekers are biased, they have $b_s = 0.5$.

the short term. Even more concerning, the last plot in Figure 3 shows that this intervention produces significantly lower chances of unprotected CCs achieving a number of followers that reflects the quality of their content, i.e., IF. This shows that interventions on recommendation and moderation to overcome biases in the viewer population are very sensitive to the chosen metric of fairness.

Seekers also benefit from bias-reducing interventions. Finally, we acknowledge that in multi-stakeholder systems, the changes that are beneficial for one party are not necessarily so for the others. Therefore, Figure 4 presents unfairness from the perspective of seekers. First and foremost, the analysis shows that, as for CCs, whenever the system is unfair for protected seekers this unfairness can be reduced through interventions that decrease existing biases in VAS or give visibility boosts to the protected group. For example, when $f = 25\%$ of seekers are biased, advantaging the protected CCs at $l = -25\%$ level leads to a fair outcome for seekers with respect to the search time needed to find the best CC (see Figure 4a). As before, different levels of compensation are needed depending on the considered metric. Second, different from CC unfairness, for any of the three analyzed metrics, if the fraction f of biased seekers decreases, then seekers will generally experience higher levels of fairness. Third, the imbalance in the population size is enough to create considerable unfairness with respect to some metrics, even when neither the VAS nor the seekers are biased (see Figure 4c). While such unfairness could be overcome by boosting the visibility of protected CCs ($l < 0$), this produces pronounced unfairness for unprotected CCs on other metrics (see Figure 4b in comparison).

6 CONCLUSION

Our work focused on understanding how audience and visibility biases contribute to inequalities on platforms with popularity-biased recommendations and moderation. Currently, content creators (CCs) and platform representatives do not agree on the prevalence of these two causes and the ways to overcome them. To aid

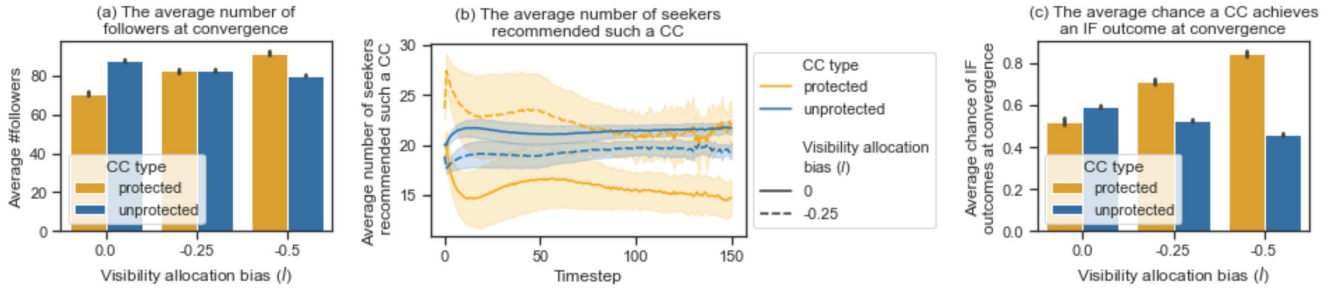

Figure 3: We consider the scenario where a fraction $f = 0.75$ of seekers are biased at a $b_s = 0.5$ level, and recommendations are made by PA. The plots show the impact of boosting the visibility of protected CCs in the VAS on (a) the average number of followers, (b) the allocated visibility over time, and (c) the average chance of achieving an IF outcome.

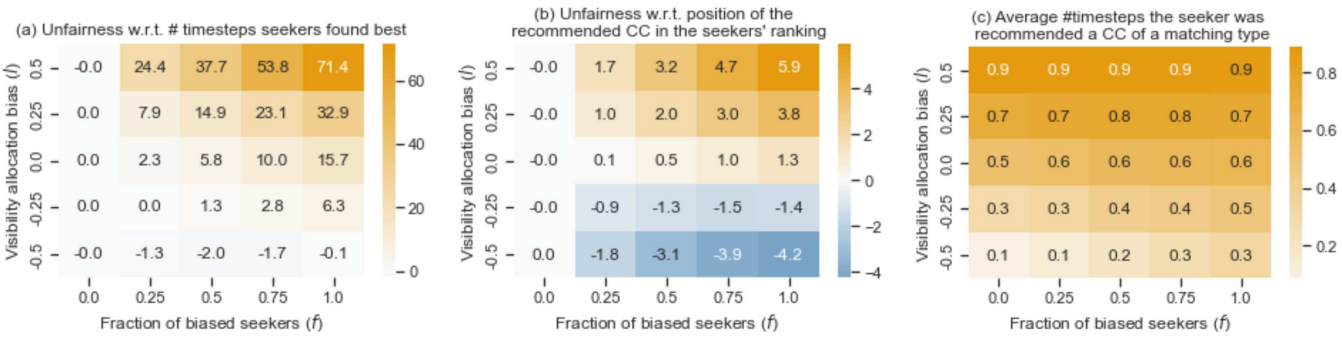

Figure 4: The plots show the impact of the fraction of homophilic biased seekers (f) and the level of visibility allocation bias (l) on the fairness of seekers with respect to (a) the number of timesteps until seekers find the best CC according to their preference, (b) the average position of the recommended CC in the ranking of the seeker, and (c) the average number of timesteps the seeker was recommended a CC of a matching type. All plots are for PA recommendations and $b_s = 0.5$.

this debate, we contributed to a better understanding of the system by building on prior literature on algorithmic fairness and defining several metrics of group fairness, which we monitored in subsequent simulations. Consistent with the beliefs of CCs, our results suggest that, when existing, biases in recommender systems and moderation pose a greater risk than the one due to the share of homophilic-biased viewers, and should thus be addressed first. However, as suggested by platforms, user biases and population imbalances could alone lead to unfairness, thus potentially making the visibility allocation seem biased. Fixing such unfairness by boosting the visibility of protected CCs is complicated: Even when it is optimally designed with respect to one metric, significant unfairness could remain or be created with respect to others. The safest solution from the platform's perspective is to decrease popularity biases: While this solution does not eliminate unfairness when users have homophilic bias, our results indicate it will generally reduce it. Finally, we investigated fairness for the audience and confirmed that interventions that improve fairness for CCs would generally also improve it for seekers.

We emphasize that our primary goal was to understand the driving mechanisms of this complex process. Thus, we leveraged a simple (but realistic [43]) model, which still allowed us to understand how inequalities in CC-centered platforms depend on the characteristics of the visibility allocation system as well as on the levels of bias within this system and within the viewer population (homophilic preferences). Despite its advantages (e.g., interpretability, allowing for causal inferences), our methodology also presents limitations and opportunities for future work. First, the model would benefit from additional data validation. While the original unidimensional model was corroborated with real datasets [43] and the extension to the multidimensional setting was, as mentioned in the text, based on prior theories, we did not directly use real-world data. Second, preserving the simplicity of the model required various simplifying assumptions. At the user level, we assume, for instance, the existence of a single binary-protected attribute, an agreement on the evaluation of same-type users, and consistency in the quality of the content created by each CC. At a platform level, examples include using non-personalized recommendations, making one recommendation per iteration, and grouping all factors that drive the formation of recommendations into a single function (i.e., the visibility allocation function). While departing from these assumptions would make our model more realistic, it would also require integrating several additional parameters. Doing so would likely compromise some of the clarity resulting from the simplicity of the model, require careful calibration of parameters to avoid modeling errors, and run into the risk of an overfit to one platform as it currently is designed, thus reducing the transferability of results. Hence, while we believe that such details would be valuable for

an in-depth analysis of specific platforms and methodologies, we opted to start with the most simple model that could still provide an overview of the most important roots of unfairness on platforms centered around CCs. Third, the simplicity of the model did not allow for comparisons of specific technical solutions to decrease biases. More precisely, when we analyze interventions, we do not investigate how efficient particular interventions are at reducing seeker biases or visibility allocation biases. Instead, we refer to the existing literature on such debiasing method [11, 17, 35] and explore how the efficacy of these methods affects inequalities in outcomes.

With these limitations noted, we believe our model and simulations already provide an initial understanding of the role played by biases of platforms and viewers on the final level of group fairness for content creators. Addressing the limitations could lead to valuable future work that enriches our understanding of CC-centered platforms; examples include extending the model to account for multiple non-binary protected attributes (e.g., both gender and race, thus allowing the study of potential issues arising at the intersection of multidimensional identities [8, 21]), evaluating more realistic recommendation algorithms together with specific technical solutions, and using suitable datasets for the calibration and evaluation of more elaborated models. Altogether, we believe this work shows the potential of leveraging cross-domain expertise (complex networks, recommender systems and moderation, prior work in psychology and sociology, agent-based modeling and simulation) in order to get insight into the long-term effects of algorithms and interventions in complex sociotechnical systems.

ACKNOWLEDGMENTS

We gratefully acknowledge the support from the University of Zürich and the SNSF (NCCR Automation grant, 180545). The authors would also like to thank the anonymous reviewers for their valuable feedback.

REFERENCES

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the eleventh ACM conference on recommender systems*. 42–46.
[2] Reed Albergotti. 2020. *Black creators sue YouTube, alleging racial discrimination.* The Washington Post. https://www.washingtonpost.com/technology/2020/06/18/black-creators-sue-youtube-alleged-race-discrimination/
[3] Greg Bensinger and Reed Albergotti. 2019. *YouTube discriminates against LGBT content by unfairly culling it, suit alleges.* The Washington Post. https://www.washingtonpost.com/technology/2019/08/14/youtube-discriminates-against-lgbt-content-by-unfairly-culling-it-suit-alleges/
[4] Clara Lindh Bergendorff. 2021. *From The Attention Economy To The Creator Economy: A Paradigm Shift.* Forbes. https://www.forbes.com/sites/claralindhbergendorff/2021/03/12/from-the-attention-economy-to-the-creator-economy-a-paradigm-shift/?sh=512e4cd8faa7
[5] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 514–524.
[6] Eszter Bokányi and Anikó Hannák. 2020. Understanding inequalities in ride-hailing services through simulations. *Scientific reports* 10, 1 (2020), 6500.
[7] Elizabeth E Bruch and MEJ Newman. 2018. Aspirational pursuit of mates in online dating markets. *Science Advances* 4, 8 (2018), eaap9815.
[8] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.
[9] Robin Burke. 2017. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093* (2017).
[10] Kyle Chayka. 2021. *What the "Creator Economy" Promises—and What It Actually Does.* The New Yorker. https://www.newyorker.com/culture/infinite-scroll/what-the-creator-economy-promises-and-what-it-actually-does
[11] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems* 41, 3 (2023), 1–39.
[12] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 525–534.
[13] We Are Social; Hootsuite; DataReportal. 2022. *Most popular social networks worldwide as of January 2022, ranked by number of monthly active users.* Statista. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/
[14] Arthur De Vany. 2003. *Hollywood economics: How extreme uncertainty shapes the film industry.* Routledge.
[15] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.
[16] Aytek Erdil and Haluk Ergin. 2008. What's the matter with tie-breaking? Improving efficiency in school choice. *American Economic Review* 98, 3 (2008), 669–689.
[17] Chloë FitzGerald, Angela Martin, Delphine Berner, and Samia Hurst. 2019. Interventions designed to reduce implicit prejudices and implicit stereotypes in real world contexts: a systematic review. *BMC psychology* 7, 1 (2019), 1–12.
[18] Werner Geyser. 2022. *Creator Earnings: Benchmark Report 2022.* Influencer Marketing Hub. https://influencermarketinghub.com/creator-earnings-benchmark-report/
[19] Werner Geyser. 2022. *The State of Influencer Marketing 2022: Benchmark Report.* Influencer Marketing Hub. https://influencermarketinghub.com/influencer-marketing-benchmark-report/
[20] Nigel Gilbert and Klaus Troitzsch. 2005. *Simulation for the social scientist.* McGraw-Hill Education (UK).
[21] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 501–512.
[22] Natali Helberger, Kari Karppinen, and Lucia D'Acunto. 2018. Exposure diversity as a design principle for recommender systems. *Information, Communication & Society* 21, 2 (2018), 191–207.
[23] Alex Hern. 2020. *TikTok 'tried to filter out videos from ugly, poor or disabled users'.* The Guardian. https://www.theguardian.com/technology/2020/mar/17/tiktok-tried-to-filter-out-videos-from-ugly-poor-or-disabled-users
[24] Influencity. 2020. *The Largest Influencer Study in Europe.* https://influencity.com/resources/studies/the-largest-influencer-study-of-europe
[25] Stefania Ionescu, Anikó Hannák, and Kenneth Joseph. 2021. An agent-based model to evaluate interventions on online dating platforms to decrease racial homogamy. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 412–423.
[26] Stefania Ionescu, Nicolò Pagan, and Anikó Hannák. 2023. Individual Fairness for Social Media Influencers. In *International Conference on Complex Networks and Their Applications*. Springer, 162–175.
[27] IZEA. 2022. *The State of Influencer Equality: 2022 Report.* https://izea.com/resources/insights/2022-state-of-influencer-equality/
[28] Jiepu Jiang, Ahmed Hassan Awadallah, Xiaolin Shi, and Ryen W White. 2015. Understanding and predicting graded search satisfaction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. 57–66.
[29] Fariba Karimi, Mathieu Génois, Claudia Wagner, Philipp Singer, and Markus Strohmaier. 2018. Homophily influences ranking of minorities in social networks. *Scientific reports* 8, 1 (2018), 11077.
[30] Matevž Kunaver and Tomaž Požrl. 2017. Diversity in recommender systems–A survey. *Knowledge-based systems* 123 (2017), 154–162.
[31] Calvin K Lai, Allison L Skinner, Erin Cooley, Sohad Murrar, Markus Brauer, Thierry Devos, Jimmy Calanchini, Y Jenny Xiao, Christina Pedram, Christopher K Marshburn, et al. 2016. Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General* 145, 8 (2016), 1001.
[32] Peggy J Liu, Brent McFerran, and Kelly L Haws. 2020. Mindful matching: Ordinal versus nominal attributes. *Journal of Marketing Research* 57, 1 (2020), 134–155.
[33] Eli Lucherini, Matthew Sun, Amy Winecoff, and Arvind Narayanan. 2021. T-RECS: A simulation tool to study the societal impact of recommender systems. *arXiv preprint arXiv:2107.08959* (2021).
[34] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 2145–2148.
[35] Dana Mastro and Riva Tukachinsky. 2012. The influence of media exposure on the formation, activation, and application of racial/ethnic stereotypes. *The international encyclopedia of media studies* (2012).
[36] Sean M McNee, John Riedl, and Joseph A Konstan. 2006. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06*

extended abstracts on Human factors in computing systems. 1097–1101.
[37] Maria Mellor. 2020. *Why is TikTok creating filter bubbles based on your race?* WIRED. https://www.wired.co.uk/article/tiktok-filter-bubbles
[38] MSL. 2021. *MSL Study Reveals Racial Pay Gap in Influencer Marketing*. https://www.mslgroup.com/whats-new-at-msl/msl-study-reveals-racial-pay-gap-influencer-marketing
[39] Goran Murić, Alexey Tregubov, Jim Blythe, Andrés Abeliuk, Divya Choudhary, Kristina Lerman, and Emilio Ferrara. 2022. Large-scale agent-based simulations of online social networks. *Autonomous Agents and Multi-Agent Systems* 36, 2 (2022), 38.
[40] Joaquim Neto, A Jorge Morais, Ramiro Gonçalves, and António Leça Coelho. 2022. Multi-agent-based recommender systems: a literature review. In *Proceedings of Sixth International Congress on Information and Communication Technology: ICICT 2021, London, Volume 1*. Springer, 543–555.
[41] Mark EJ Newman. 2001. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences* 98, 2 (2001), 404–409.
[42] Jakob Nielsen. 2006. Participation inequality: lurkers vs. contributors in internet communities. *Jakob Nielsen's Alertbox* 107 (2006), 108.
[43] Nicolò Pagan, Wenjun Mei, Cheng Li, and Florian Dörfler. 2021. A meritocratic network formation model for the rise of social media influencers. *Nature communications* 12, 1 (2021), 1–12.
[44] Amy Pei, Yakov Bart, Koen Pauwels, and Kwong Chan. 2022. Racial Pay Gap in Influencer Marketing. *Available at SSRN 4156872* (2022).
[45] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *science* 311, 5762 (2006), 854–856.
[46] Thomas C. Schelling. 1971. Dynamic models of segregation. *The Journal of Mathematical Sociology* 1, 2 (1971), 143–186. https://doi.org/10.1080/0022250X.1971.9989794
[47] Stephen A Spiller and Lena Belogolova. 2017. On consumer beliefs about quality and taste. *Journal of Consumer Research* 43, 6 (2017), 970–991.
[48] Chandra Steele. 2022. *How Racial Inequalities Affect Influencers*. PCMag. https://influencity.com/resources/studies/the-largest-influencer-study-of-europe
[49] Louise T Su. 2003. A comprehensive and systematic model of user evaluation of web search engines: I. Theory and background. *Journal of the American society for information science and technology* 54, 13 (2003), 1175–1192.
[50] Morgan Sung. 2019. *TikTok users of color call for better visibility on the For You Page*. Mashable. https://mashable.com/article/tiktok-users-of-color-call-for-visibility-for-you-page
[51] Douglas R Turnbull, Sean McQuillan, Vera Crabtree, John Hunter, and Sunny Zhang. 2022. Exploring Popularity Bias in Music Recommendation Models and Commercial Steaming Services. *arXiv preprint arXiv:2208.09517* (2022).
[52] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. *Advances in neural information processing systems* 30 (2017).
[53] Jingjing Zhang, Gediminas Adomavicius, Alok Gupta, and Wolfgang Ketter. 2020. Consumption and performance: Understanding longitudinal dynamics of recommender systems via an agent-based simulation framework. *Information Systems Research* 31, 1 (2020), 76–101.
[54] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 11–20.