Algorithms that "Don't See Color": Measuring Biases in Lookalike and Special Ad Audiences

Piotr Sapiezynski
Northeastern University
Boston, MA, USA
p.sapiezynski@northeastern.edu

Avijit Ghosh
Northeastern University
Boston, MA, USA
ghosh.a@northeastern.edu

Levi Kaplan
Northeastern University
Boston, MA, USA
kaplan.l@northeastern.edu

Aaron Rieke
Upturn
Washington, DC, USA
aaron@upturn.org

Alan Mislove
Northeastern University
Boston, MA, USA
amislove@ccs.neu.edu

ABSTRACT

Researchers and journalists have repeatedly shown that algorithms commonly used in domains such as credit, employment, healthcare, or criminal justice can have discriminatory effects. Some organizations have tried to mitigate these effects by simply removing sensitive features from an algorithm's inputs. In this paper, we explore the limits of this approach using a unique opportunity. In 2019, Facebook agreed to settle a lawsuit by removing certain sensitive features from inputs of an algorithm that identifies users similar to those provided by an advertiser for ad targeting, making both the modified and unmodified versions of the algorithm available to advertisers. We develop methodologies to measure biases along the lines of gender, age, and race in the audiences created by this modified algorithm, relative to the unmodified one. Our results provide experimental proof that merely removing demographic features from a real-world algorithmic system's inputs can fail to prevent biased outputs. As a result, organizations using algorithms to help mediate access to important life opportunities should consider other approaches to mitigating discriminatory effects.

CCS CONCEPTS

• **Social and professional topics** → **Computing / technology policy**; • **Applied computing** → *Marketing*.

KEYWORDS

online advertising, fairness, process fairness

ACM Reference Format:
Piotr Sapiezynski, Avijit Ghosh, Levi Kaplan, Aaron Rieke, and Alan Mislove. 2022. Algorithms that "Don't See Color": Measuring Biases in Lookalike and Special Ad Audiences. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3514094.3534135

AIES'22, August 1–3, 2022, Oxford, United Kingdom

ACM ISBN 978-1-4503-9247-1/22/08…$15.00
https://doi.org/10.1145/3514094.3534135

1 INTRODUCTION

Organizations use algorithmic models[1] ("algorithms") in a variety of important domains, including healthcare [27], credit [19], employment [9, 23], and content distribution [3]. Unfortunately, these algorithms have been shown to sometimes have discriminatory effects that can often be challenging to detect, measure, and articulate. Some have proposed mitigating discriminatory effects by removing demographic features from an algorithm's inputs. For example, in 2019 the U.S. Department of Housing and Urban Development (HUD) proposed a rule that considered applying this approach to housing discrimination [12]. Because algorithms can effectively use omitted demographic features by combining other inputs that are each *correlated* with those features [5], such a rule could nullify any protection from discriminatory effects. This is particularly true in large-scale machine learning (ML) systems, which can take as input thousands or even millions of features [6].

In this paper, we leverage a unique opportunity created by a recent lawsuit settlement involving Facebook's advertising platform to explore the limits of this approach. Specifically, we examine Facebook's *Lookalike Audiences* targeting tool, which takes a list of Facebook users provided by an advertiser (called the *source audience*) and creates a new audience of users who share "common qualities" with those in the source audience. In March 2018, the National Fair Housing Alliance (NFHA) and others sued [13] Facebook over violations of the Fair Housing Act (FHA). When the case was settled in March 2019, Facebook agreed to modify the functionality of Lookalike Audiences when used to target housing, credit, and employment ads. In brief, Facebook created the *Special Ad Audiences* tool, which works like Lookalike Audiences, except its algorithm does *not* consider users' age, gender, relationship status, religious views, school, political views, interests, or zip code when detecting common qualities.

We seek to learn whether the Special Ad Audience algorithm (which is not provided with certain demographic features) actually produces significantly less *skewed* audiences than the Lookalike Audience algorithm (which is). In other words, when provided with a source audience that skews heavily toward one demographic group over another, to what extent do each of these tools reproduce that skew? We focus on skews along demographic features named

[1] Throughout this paper, we refer to a large class of algorithmic models using the now-common term "algorithms", especially those created through statistical modeling and machine learning.

in the settlement, enabling us to examine whether simply removing the protected features as input to an algorithm is sufficient eliminate skew along those features. To do so, we develop a methodology to examine the delivery of the same ads when using the two types of audiences, measuring the skew along the lines of gender, age, and race.

We show that our Special Ad audiences[2] are skewed to almost the same degree as Lookalike audiences, with many of the results being statistically indistinguishable. For example, when using a source audience that is all women, our Lookalike audience-targeted ad delivered to 96.1% women, while Special Ad audience-targeted ad delivered to 91.2% women. We also provide evidence indicating that both Lookalike and Special Ad audiences carry—to a certain extent—the biases of the source audience in terms of race and political affiliation.

To underscore the real-world impact of these results, we place ads as an employer who is seeking to find candidates "similar to" to their current workforce. Using a source audience consisting of Facebook employees we find that the resulting Special Ad audience skews heavily towards 25–34-year-old men. We also confirm that previous findings on how Facebook's delivery mechanisms can cause *further* skews in who is shown ads hold for Special Ad Audiences.

Taken together, our results show that simply removing demographic features from the inputs of a large-scale, real-world algorithm will not always suffice to meaningfully change its outputs. At the same time, this work presents a methodology by which other algorithms could be studied.

To be clear, we are not claiming—and do not believe—that Facebook has *incorrectly* implemented Special Ad Audiences, or is in violation of its settlement agreement. Rather, the findings in this paper are a natural result of how complex algorithmic systems work in practice.

Ethics The research has been reviewed by our Institutional Review Board and marked as exempt. Further, we minimized harm to Facebook users by only running "real" ads, i.e., if a user clicked on one of our ads, they were presented with a real-world site relevant to content the ad. We did not have any direct interaction with the users who were shown our ad, and did not collect any of their information. Finally, we minimized harm to Facebook by running and paying for our ads just like any other advertiser, as well as flagging them as employment ads whenever applicable.

2 BACKGROUND

In this section, we provide background on Facebook's ad targeting tools and overview related work.

2.1 Facebook's ad targeting tools

Facebook provides a range of *targeting* tools to help advertisers select an *audience* of users who will be eligible to see their ads. For example, advertisers can select users through combinations of *targeting attributes*, including over 1,000 demographic, behavioral, and interest-based features.

[2]Throughout the paper, we use "Lookalike Audience" or "Special Ad Audience" to refer to the general tools provided by Facebook, and "Lookalike audience" or "Special Ad audience" to refer to a particular audience.

More germane to this paper and its methods, Facebook also offers a number of other, more advanced targeting tools. One such tool is *Custom Audiences*, which allows advertisers indicate individual users that they wish to include in an audience. To use Custom Audiences, an advertiser uploads a list of personally identifiable information (PII), potentially including names, email addresses, phone numbers, dates of birth, and mobile identifiers. Facebook then compares those identifiers against its database of active users, and lets the advertiser include matched users in their target audience.

Another tool is *Lookalike Audiences*, which creates an audience of users who share "common qualities" with users in a Custom audience provided by the advertiser (called the *source audience*). The exact input qualities used by the algorithm in creating these audiences are not known and the documentation lists only two examples: demographic information and interests. Prior work has demonstrated that Lookalike Audiences can reproduce demographic skews present in source audiences [34].

2.2 Special Ad Audiences

In March 2018, the NFHA and others sued Facebook for allowing landlords and real estate brokers to exclude members of protected groups from receiving housing ads [13]. The lawsuit was settled in March 2019, and Facebook agreed to make a number of changes to its ad targeting tools. Facebook now refers to this modified Lookalike Audiences tool as *Special Ad Audiences*.

From an advertiser's perspective, Special Ad Audiences are created in the same manner as Lookalike Audiences (i.e., based on a source Custom audience). The minimum size for both types of these algorithmically generated audiences is 1% of the population of the target location, regardless of the size of the source audience. In case of the US that means that the algorithm outputs audiences of 2.3 million users.

2.3 Related work

Greenberg distinguishes two kinds of fairness concerns, *distributive* and *procedural* [22]. The former aims to assure balanced outcomes, whereas the latter focuses on the process itself. Elimination of sensitive features, for example sex or race, from an algorithm's input (as with Special Ad Audiences) falls into the *procedural* category. Such approach in the legal context is also referred to as *anti-classification* and it is encoded in the current standards [11]. However, scholars and researchers have for decades critiqued this so-called "colorblind" approach to addressing historical inequality and discrimination [7]. Legal scholar Destiny Perry argues that "(1) colorblindness is, under most circumstances, undesirable given its recently discovered negative outcomes, particularly for the very groups or individuals it is meant to protect; (2) true colorblindness is unrealistic given the psychological salience of race; and (3) race consciousness in the law is necessary to ensure equal treatment of racial groups in regulated domains such as housing, education, and employment [30]." In the context of sentencing and mass incarceration Traci Schlesinger concludes that "in the post-civil rights era, racial disparities are primarily produced and maintained by colorblind policies and practices [32]." Similar arguments have been made in the context of housing discrimination and a range of other domains [4].

Previous work in statistics and machine learning indicated that, in general, removing sensitive features does not reliably achieve fairness for a number of reasons. First, certain features might serve as close proxies for the sensitive information. For example, due to housing segregation a person's zip-code can be predictive of their race. Second, the removed information might be *redundantly encoded* by non-sensitive features or their combinations. It will then be reconstructed by the model if it is pertinent to the prediction task [10, 14, 36]. One such example is the fiasco of Amazon's hiring algorithm [21]. Third, there are cases in which only certain intersections of values of otherwise non-sensitive features are to be protected [29]. Finally, even if none of the features or their combinations are unfair, their predictive performance might differ across sub-populations. In an effort to minimize the total error, the classifier will fit the majority group better than the minority [8, 31]. Taken together, these prior works paint a clear picture of process fairness, or fairness through unawareness, as insufficient to ensure fair outcomes. Unfortunately, despite this consensus among scholars and a few high-profile failures in practice, the 2019 settlement is still based on fairness through unawareness. In this article we investigate whether this particular implementation is closer to achieving the goal of fairness.

Regardless of the particular approach to ML fairness, focusing on particular algorithms can be too narrow of a problem definition. Real-world algorithmic systems are often composed of multiple subsystems and can be discriminatory as a whole, even if built from a series of fair algorithms [15]. They need to be modeled along with the other components of the *socio-technical* systems they are embedded in [33]. The burden of these investigations lies on independent researchers and auditors since the companies who operate these algorithms might not be incentivized to measure and address the externalities they cause [28].

3 METHODOLOGY

In this work we attempt to measure the audience skews in terms of gender, age, race, and political views. Facebook Ad Manager reports the gender and age distribution of the audiences that received each ad, but it does not report the information about the race or political views of these audiences. We therefore apply two different approaches to creating the audiences and measuring the effects.

3.1 Timing

The 2019 settlement [16] stipulated that the updated ad creation flow for special categories be implemented by September 30, 2019. All of our ads were created and run between October 20, 2019 and December 15, 2019, leaving Facebook ample time after the implementation deadline.

3.2 Measuring skews by gender and age

To measure the makeup of a target audience by gender and age, we create and run actual ads and then we use the Facebook Ad Manager API to record how they are delivered. For these experiments, we need to provide an *ad creative* (consisting of the ad text, headline, image, and destination URL). Since the ad content influences the delivery [3], we chose to use the same creative for all ads, unless otherwise noted: a generic ad for Google Web Search, which has basic text ("Search the web for information") and a link to Google. We found that Facebook does not verify that an ad that is self-reported by an advertiser as a housing, credit, or employment ad is, in fact, such an ad. On the other hand, Facebook does automatically classify housing, credit, or employment ads as such even if the advertisers chooses not to disclose that information. Thus, the only way for us to run the same ad creative using both Lookalike and Special Ad audiences was to run a neutral ad that would not trigger the automatic classification.

Creating audiences Recall that our goal is to measure whether Special Ad Audiences produce significantly less biased audiences than Lookalike Audiences. We therefore need to generate source audiences with controlled and known bias, from which we can create a Lookalike and a Special Ad audience. We replicate the approach from prior work [3], relying on publicly available voter records from New York and North Carolina. These records include registered voters' gender, age, location (address), and (only in North Carolina) race.

Thus, for each demographic feature we wish to study, we first create a Custom audience based on the voter records (which we treat as ground truth). For example, when studying gender, we select a subset of the voters who are listed as female and use that list to create a Custom audience. We use each biased Custom audience to create both a Lookalike audience and a Special Ad audience, selecting users in the U.S. and choosing the smallest size option (1% of the population).

Data collection Once the ads are running we use Facebook's Ad Manager tool to collect information about demographics of the audiences that Facebook shows our ads to, broken down by age group, gender, and the intersections of these two characteristics.

Calculating and comparing gender skew The Ad Manager tool reports gender of each user as either female, male, or unknown. The unknown gender might refer to users who choose to self-report their gender as falling outside of the binary, or those who did not provide their gender. We note that in all experiments there is no more than 1% of such users, and report the observed gender bias as the fraction of men $\hat{p}$ in the reached audience. We also calculate the upper and lower 99% confidence intervals ($U.L$ and $L.L$, respectively) around this fraction $\hat{p}$ using the method presented by Agresti and Coull [2]:

$$\begin{aligned} L.L. &= \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + z^2_{\alpha/2}/n}, \\ U.L. &= \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + z^2_{\alpha/2}/n}, \end{aligned} \tag{1}$$

We set $z_{\alpha/2} = 2.576$, corresponding to the 99% interval.

Finally, we verify whether the difference between fractions observed for Lookalike and Special Audiences is statistically significant using the difference of proportion test:

$$\Delta_{p_{LS}} = (\hat{p_L} - \hat{p_S}) \pm z_{\alpha/2}\sqrt{\frac{\hat{p_L}(1-\hat{p_L})}{n_L} + \frac{\hat{p_S}(1-\hat{p_S})}{n_S}}, \tag{2}$$

where $\hat{p_L}$ and $\hat{p_S}$ are the fractions of men who saw the ad in the Lookalike and Special audiences, n_L and n_S are number of people reached in each of these audiences. Because we are testing the significance in seven experiments (one for each input proportion), we apply the Bonferroni correction for multiple hypotheses testing. We do so by setting $z_{\alpha/2}$ to 3.189, corresponding to Bonferroni corrected $p_{val} = 0.01/7 \approx 0.00143$. If the confidence interval includes 0, we cannot reject the hypothesis that the fraction of men is the same in the two audiences and thus the result is not statistically significant.

Calculating and comparing the age skew Age of the users who were shown each ad is reported in groups: <18, 18-24, 35-44, 45-54, 55-64, and 65+. We calculate the mean age and the confidence intervals around it using formulas specific to grouped data. First, we compute the mid-point M_i for each age range i,

$$M_i = \frac{x_{min_i} + x_{max_i}}{2} \tag{3}$$

Next, we find the mean age μ

$$\mu = \frac{\sum_i (M_i \times F_i)}{\sum_i F_i}, \tag{4}$$

where F_i is the number of audience members in the age group i. We then compute the standard deviation around that mean

$$\sigma = \sqrt{\frac{\sum_i (F_i \times M_i^2) - (n \times \mu^2)}{n-1}} \tag{5}$$

and the corresponding standard error

$$SE = \frac{\sigma}{\sqrt{n}} \tag{6}$$

Presented upper and lower confidence intervals correspond to

$$\begin{aligned} U.L. &= \mu + k \times SE, \\ L.L. &= \mu - k \times SE \end{aligned} \tag{7}$$

respectively, and k is set to 2.576.

Finally, we verify whether the difference in mean ages between the Lookalike and Special audiences is statistically significant. To achieve that, we compute the standard error of the difference

$$SE_{LS} = \sqrt{\frac{\sigma_L^2}{n_L} + \frac{\sigma_S^2}{n_S}} \tag{8}$$

and the 99% confidence interval around the difference between mean ages:

$$\Delta_{\mu_{LS}} = \mu_L - \mu_s \pm z_{\alpha/2} \times \sqrt{\frac{\sigma_L^2}{n_L} + \frac{\sigma_S^2}{n_S}} \tag{9}$$

We apply the Bonferroni correction for six tests and use the $z_{\alpha/2}$ set to 3.143. If the confidence interval includes 0, we cannot reject the hypothesis that the mean age is the same in the two audiences and thus the difference is not statistically significant.

3.3 Measuring racial skews

When measuring racial skew in the audiences we are unable to re-use the same methodology for age and gender, which relied on Facebook's *ad delivery* statistics. Instead, we develop an alternative methodology that relies on *estimated daily results*– Facebook's estimate of the number of users matching the advertiser's targeting criteria that can be reached daily within the specified budget. We set the daily budget to the maximum allowed value ($1M) to best approximate the total number of users that match the targeting criteria. Facebook returns these values as a range (e.g., "12,100 – 20,400 users"); throughout this procedure, we always use the lower value.[3] The procedure has only two steps: audience creation and targeting. It does not involve running any ads and observing the skew in delivery, and it is entirely based on the estimates on audience sizes provided by Facebook at the ad targeting step.

We note that ours is not the first use of these estimates to infer the number of users that match different criteria. For example, Garcia et al. used them to estimate the gender inequality across the globe [20], while Fatehkia et al. found they are highly predictive of a range of other social indicators [18].

Audience Creation We start with the publicly available voter records from North Carolina, in which the voters self-report their race and ethnicity. We focus on two groups: Non-Hispanic Black and Non-Hispanic white. For each group, we create two independent Custom audiences: one list of 10,000 randomly selected users with that race, and one list of 900,000 randomly selected users with that race. The latter audience does not contain any individuals already selected for the first list, and will be refered to as the *reference* audience.

We refer to these as `w_10k` and `w_900k` (white audiences) and `b_10k` and `b_900k` (Black audiences). We then have Facebook algorithmically generate Lookalike and Special Ad audiences using the smaller Custom audiences as input. We refer to the resulting audiences as $L_{\texttt{w\_10k}}$ (for the Lookalike audience based on `w_10k`), $S_{\texttt{w\_10k}}$ (for the Special Ad audience), $L_{\texttt{b\_10k}}$, and $S_{\texttt{b\_10k}}$.

Targeting The goal of this step is to find the overlaps between the audiences with unknown race generated by the algorithms and the reference Custom audiences that we provided (with known race). Then we can say there is a race bias in the white Lookalike audience $L_{\texttt{w\_10k}}$ if the overlap between it and a white reference audience `w_900k` is higher than the overlap between it and a Black reference audience `b_900k` (and vice versa for an audience generated from a Black source audience). We also perform these overlap comparisons for Special Ad audiences to measure whether this effect persists despite removing sensitive features from the algorithm.

Our method relies on the fact that Facebook allows advertisers not only to specify which audiences to *include* in the targeting, but also which to *exclude*. Suppose we wish to obtain an estimate of the fraction of white users in $L_{\texttt{w\_10k}}$. To do so, we first target the reference white audience `w_900k` audience and record the potential daily reach (e.g., 81,000). We then target $L_{\texttt{w\_10k}}$ and record the potential daily reach (e.g., 397,000). Finally, we target $L_{\texttt{w\_10k}}$ and *exclude* the `w_900k` audience, and record the potential daily reach (e.g., 360,000). Now, we can observe that excluding `w_900k` from

[3] We used the midpoint and the upper value and found similar results.

L_{w_10k} caused the potential daily reach to drop by 37,000, indicating that approximately 46% (37,000/81,000) of w_900k were present in L_{w_10k}. We can then repeat the process with excluding b_900k, and measure the fraction of the reference Black audience that is present in L_{w_10k}. By comparing the fraction of w_900k and b_900k that are present in L_{w_10k}, we obtain an estimate of the racial bias of L_{w_10k}.

Measuring political skews To measure political skews we follow the exact same method as with measuring racial skews, but rather than constructing the audiences based on their reported race, we use their registered political affiliation as Democratic or Republican voters.

Limitations Unlike in our experiments with gender and age, here we do not know the race of a vast majority of the audience. The Lookalike and Special Ad audiences that Facebook creates consist mostly of people who appear not to be in our voter records. There are multiple reasons for why this might be the case: (1) we only looked and single race, non-Hispanic white and Black voters, excluding all Hispanic voters, as well as those of other races, and multi-racial; (2) the users in the created audiences and could be located in other states - while creating lookalike and special audiences the advertiser can only select the country where those audiences would be located. Thus, the results we present in this section only refer to the fraction of voters with known race who are included in each Lookalike and Special Ad audience, not the racial composition of these audiences overall. Still, these estimates do give us a small window into the makeup of the Lookalike and Special Ad audiences.

4 RESULTS

We now present our experiments and analyze whether Lookalike and Special Ad Audiences show similar levels of skew.

4.1 Gender and age

We begin by focusing on gender, creating seven Custom audiences based on New York voter records. Each audience contains 10,000 voters, with varying fractions of men: 0%, 20%, 40%, 50%, 60%, 80%, 100%. We run ads to the resulting Lookalike and Special Ad audiences, and compare the results in ad delivery as reported by Facebook's advertiser interface.

Figure 1A presents a summary of the results of this experiment, and we make a number of observations. *First*, we can see that each Lookalike audience clearly mirrors its source audience along gender lines: the Lookalike audience derived from a male-only source audience delivers to over 99% men, and the the Lookalike audience derived from a female-only source audience delivers to over 97% women. *Second*, we observe a slight male bias in our delivery, relative to the source audience: for example, the Lookalike audience derived from a source audience of 50% men actually delivered to approximately 70% men. This male bias has been observed by prior work [3, 26] and may be due to market effects or ad delivery effects (which affect both Lookalike and Special Ad audiences equally). *Third*, and most importantly, when we compare the delivery of each Special Ad audience to its corresponding Lookalike audience,

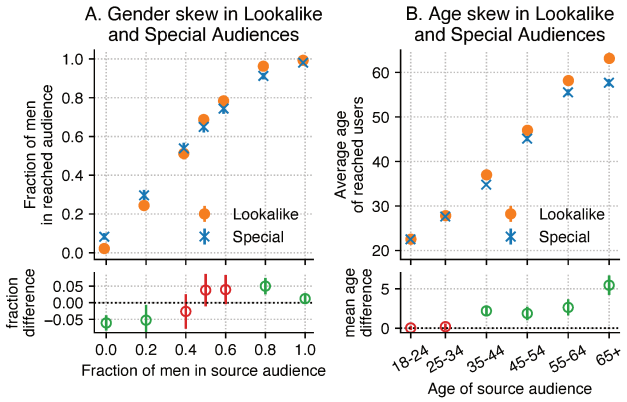

Figure 1: A. Gender breakdown of ad delivery to Lookalike and Special Ad audiences created from the source audiences with varying fraction of male users. The Special Ad audiences replicate the skew to a large extent. B. Age breakdown of ad delivery to Lookalike and Special Ad audiences created from source audiences with varying age brackets. Both Lookalike and Special Ad audiences follow the age distribution of the source audiences, but the latter shows a decrease of mean age by up to six years in the 65+ group.

we observe that a similar level of skew (that in some cases is statistically indistinguishable). For example, the Special Ad audience derived from a male-only source audiences delivers to over 95% men, despite being created without having access to users' genders. As emphasized the lower panel of Figure 1, the Special Ad audiences do show a bit less skew when compared to the Lookalike audiences for some of the input audiences, while still carrying over most of the skew from the source audience.

We follow an analogous procedure to create six Custom Audiences, each consisting of individuals only in a specified age range. We then create Custom and Special Ad audiences and measure whether the age skews are reproduced and present the results in Figure 1B.

4.2 Race

Next, we turn to examine the extent to which Special Ad Audiences can be biased along racial lines, in the same manner Lookalike Audiences were observed to be in past work [34]. We summarize the overlap between the Lookalike and Special Ad audiences and the large white and Black audiences in Table 1. Focusing on the table, we can immediately observe that both the Lookalike audiences show significantly more overlap with the race of the source audience, suggesting that the makeup of the Lookalike audiences are racially biased. For example, the Lookalike audience created from b_10k contains 61% of the active users from b_900k but only 16% of the active users from w_900k (see Methods for the explanation of the audience names). More importantly, the Special Ad audiences show a similar behavior (though as before, perhaps with slightly less of a bias). Again, it is important to keep in mind that we can only make estimates of the fraction of w_900k and b_900k that overlap with the Lookalike and Special Ad audiences, and cannot comment on the majority of these audiences (as they likely fall outside of

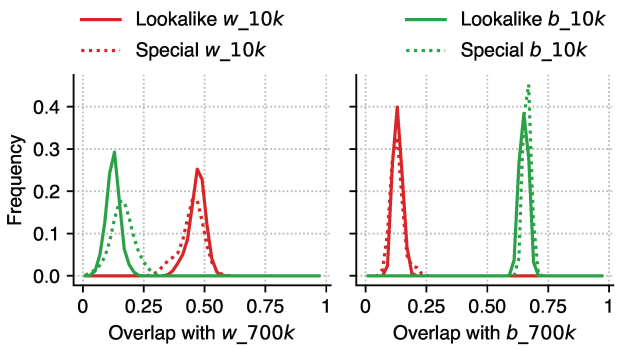

Figure 2: Both Lookalike and Special Ad audiences created from source audiences of white users containing a higher fraction of white users than Black users. Conversely, audiences created from source audiences of Black users contain a higher fraction of Black users than white users.

North Carolina). Thus, our results are not conclusive—but only suggestive—that the overall audiences are similarly biased. Below, we provide further robustness analysis of these results.

4.3 Robustness

Here, we verify that the presented results regarding race biases are robust to the random selection of seed from which Lookalike and Special Ad audiences are created. Following the method described in Methodology, we use the two sample Kolmogorov–Smirnov test to compare the distributions of overlaps presented in Figure 2. The findings are confirmed to be robust to the particular source audience choice. *First*, the racial skew observable in Lookalike audiences persists in Special Ad audiences and is statistically significant at $p_{val} = 0.01$ even with the Bonferroni correction for multiple hypotheses testing, *Second*, the differences between overlaps produced by Special Ad audiences and Lookalike audiences generated from the `w_10k` custom audience *are not statistically significant* – Special Ad audiences generated from the `w_10k` are just as biased as the corresponding Lookalike audiences. *Third* the differences between overlaps produced by Special Ad audiences and Lookalike audiences generated from the `b_10k` custom audience *are small statistically significant* and this difference comes from Special Ad audiences being even more biased than Lookalike audiences.

4.4 Political views

We next turn to measure the extent to which Lookalike and Special Ad Audiences can be biased along the lines of political views. As with race, Facebook does not provide a breakdown of ad delivery by users' political views. Thus, we repeat the methodology we used for race, using voter records from North Carolina and focusing on the differences in delivery to users registered as Republicans and Democrats.

We report the results in Table 2. We can observe a skew along political views for Lookalike audiences (for example, the Lookalike audience created from users registered as Democrats contains 51% of `d_900k` but only 32% of `r_900k`). We can also observe that the Special Ad audiences show a skew as well, though to a somewhat lesser degree than the Lookalike audiences. As with the race experiments, we remind the reader that we can only observe the overlap between the created audiences and the large Democrat/Republican audiences; we are unable to measure the majority of the created audiences. However, the demonstrated skew suggests that there is a bias in the overall makeup of the created audiences.

4.5 Real-world use cases

Next, we test a "real-world" use case of Special Ad Audiences. We imagine an employer wants to use Facebook to advertise open positions to people who are similar to those already working for them. The employer might assume that since the Special Ad Audiences algorithm is not provided with protected features as inputs, it will allow them to reach users who are similar to their current employees but without gender, age, or racial biases. The employer would therefore upload a list of their current employees to create a Custom audience, ask Facebook to create a Special Ad audience from that, and then target job ads to the resulting Special Ad audience.

We play the role of this hypothetical employer (Facebook itself in this example, which provides employees with an `@fb.com` email address). We then run the following experiment: We first create a baseline audience by using randomly generated U.S. phone numbers, 11,000 of which Facebook matched to existing users. We then create a Custom audience consisting of 12M generated email addresses: all 2–5 letter combinations + `@fb.com`, 11,000 of which Facebook matched to existing users; this is our audience of Facebook employees. We create Special Ad audiences based on each of

Source	Type	Percent overlap Black (`b_900k`)	Percent overlap White (`w_900k`)
100% Black	Lookalike ($L_{\texttt{b\_10k}}$)	61.0	16.0
	Special ($S_{\texttt{b\_10k}}$)	62.3	12.3
100% white	Lookalike ($L_{\texttt{w\_10k}}$)	16.9	42.0
	Special ($S_{\texttt{w\_10k}}$)	10.4	35.8

Table 1: Breakdown of overlap between audiences with known racial makeup and Lookalike and Special Ad audiences. While we do not know the race of the vast majority of the created audiences, we see large discrepancies in the race distribution among the known users.

Source	Type	Percent overlap Democrat (`d_900k`)	Percent overlap Republican (`r_900k`)
Democrats	Lookalike $L_{\texttt{d\_10k}}$	51.6	31.8
	Special $S_{\texttt{d\_10k}}$	42.2	25.8
Republicans	Lookalike $L_{\texttt{r\_10k}}$	28.1	50.0
	Special $S_{\texttt{r\_10k}}$	25.0	47.0

Table 2: Breakdown of overlap between source audiences with known political leaning and resulting Lookalike and Special Ad audiences. While we do not know the political leaning of the vast majority of the audiences, we see discrepancies in the distribution among the known users.

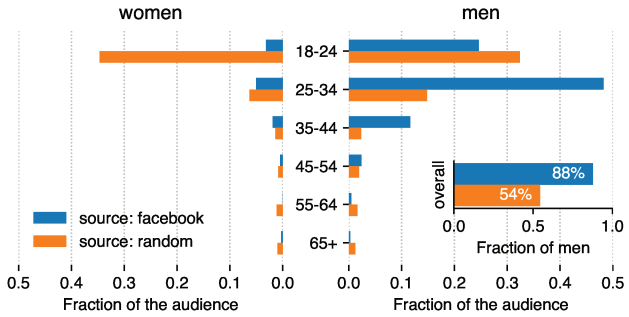

Figure 3: Gender and age breakdown of a generic job ad delivery to a Special Ad audience based on random American users (in orange) and a Special Ad audience based on Facebook employees (in blue). The audience based on Facebook employees is predominantly male and 25-34.

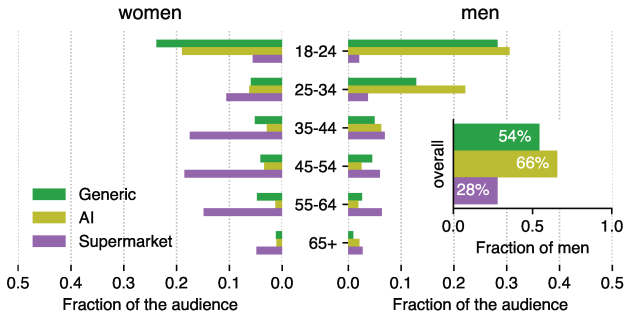

Figure 4: Gender and age breakdown of delivery of job ads to a Special Ad audience based on random American users. Facebook's delivery optimization based on the ad content can lead to large skews despite the gender and age-balanced target audience.

these two Custom audiences. Finally, we run two generic job ads —each to one of these Special Ad audiences, at the same time, from the same account, with the same budget—and observe how they are delivered.

Figure 3 presents the results of the experiment. The Special Ad audience based on Facebook employees delivers to 88% men, compared to 54% in the baseline case. Further, the Special Ad audience based on Facebook employees delivers to 48% to men aged between 25-34, compared to 15% for the baseline audience. Note that Facebook themselves report that the actual skew among company employees is lower, with 63% of male employees [1]. Overall, our results show that our hypothetical employer's reliance on Special Ad audiences to avoid discrimination along protected classes was misplaced: their ad was ultimately delivered to an audience that was significantly biased along age and gender lines (and presumably reflective of Facebook's employee population). Based on this singular experiment we cannot claim that the extent of the problem would be similar for other employers. Still, we do recommend that potential advertisers use the tool cautiously.

4.6 Content-based skew in delivery

Previous work [3, 24] demonstrated that the skew in delivery can be driven by Facebook's estimated relevance of a particular ad copy to a particular group of people. Specifically, even when the target audience were held constant, Facebook would deliver our ads to different subpopulations: ads for supermarket jobs were shown primarily to women, while ads for jobs in lumber industry were presented mostly to men. Here, we show that these effects persist also when using Special Ad Audiences. We run generic job ad to a Special Ad Audience created from a random set of 11,000 users along with ads for supermarket and artificial intelligence pointing to search for either keyword on `indeed.com`. Figure 4 shows that the different ads skew towards middle-aged women (in the case of supermarket jobs) or towards younger men (in the case of artificial intelligence jobs).

The results underline a crucial point: when designing fairness/anti-discrimination controls, one cannot just focus on one part of the *algorithmic* system. Instead one must look at the whole *socio-technical* system, including how an algorithm is used by real people, how people adjust their behaviors in response to the algorithm, and how the algorithm adapts to people's behaviors.

5 LEGAL IMPLICATIONS

At a high level, U.S. federal law prohibits discrimination in the marketing of housing, employment and credit opportunities. Our findings might have near-term legal consequences for advertisers and even Facebook itself.

A creditor, employer, or housing provider who used biased Special Ad audiences in their marketing could run afoul of the US anti-discrimination laws. This could be exceptionally frustrating for an advertiser who believed that Special Ad Audiences was an appropriate, legally-compliant way to target their ads.

Facebook itself could also face legal scrutiny. In the U.S., Section 230 of the Communications Act of 1934 (as amended by the Communications Decency Act, specifically 47 USC § 230 *Protection for private blocking and screening of offensive material*) provides broad legal immunity to Internet platforms acting as publishers of third-party content. This immunity was a central issue in the litigation resulting in the settlement analyzed above. Although Facebook argued in court that advertisers are "wholly responsible for deciding where, how, and when to publish their ads" [17], this paper makes clear that Facebook can play a significant, opaque role by creating biased Lookalike and Special Ad audiences. If a court found that the operation of these tools constituted a "material contribution" to illegal conduct, Facebook's ad platform could lose its immunity [35].

6 DISCUSSION

We demonstrated that both Lookalike and Special Ad Audiences can create similarly biased target audiences from the same source audiences. We are not claiming that Facebook incorrectly implemented Special Ad Audiences, nor are we suggesting they violated the settlement. Rather, our findings are a consequence of a complex algorithmic system at work.

Our findings have broad and narrow implications. Broadly, we demonstrate that simply removing demographic features from a complex algorithmic system can be insufficient to remove bias from its outputs, which is an important lesson for government and corporate policymakers. More specifically, we show that relative to Lookalike Audiences, Facebook's Special Ad Audiences do little to reduce demographic biases in target audiences. As a result, we believe Special Ad Audiences will do little to mitigate discriminatory outcomes.

Absent any readily available algorithm-centered solutions to the presented problem, removing the Lookalike/Special Ad audience functionality as well as disabling ad delivery optimization in the sensitive contexts of housing, employment, and credit ads might be the appropriate interim approach.

ACKNOWLEDGEMENTS

The authors thank Ava Kofman and Ariana Tobin for suggesting the experiments presented in Section 4.5 as well as for going an extra mile (or two) for their ProPublica story around this work [25]. We also thank NaLette Brodnax for her feedback on the experimental design and Aleksandra Korolova for her comments on the manuscript. This work was funded in part by a grant from the Data Transparency Lab, NSF grants CNS-1916020 and CNS-1616234, and Mozilla Research Grant 2019H1.

REFERENCES

[1] Advancing Opportunity For All. https://diversity.fb.com/read-report/.
[2] Alan Agresti and Brent A Coull. Approximate Is Better Than "exact" For Interval Estimation Of Binomial Proportions. *The American Statistician*, 52(2):119–126, 1998.
[3] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Discrimination Through Optimization: How Facebook's Ad Delivery Can Lead To Biased Outcomes. In *ACM Conference on Computer Supported Cooperative Work*, Austin, Texas, USA, November 2019.
[4] Michelle Wilde Anderson. Colorblind segregation: Equal protection as a bar to neighborhood integration. *Calif. L. Rev.*, 92:841, 2004.
[5] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness And Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.
[6] Joseph Blass. Algorithmic Advertising Discrimination. *Northwestern University Law Review*, 114(2):415–468, 2019.
[7] Eduardo Bonilla-Silva. *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States*. Rowman & Littlefield Publishers, 2006.
[8] Irene Chen, Fredrik D Johansson, and David Sontag. Why Is My Classifier Discriminatory? In *Advances in Neural Information Processing Systems*, pages 3539–3550, 2018.
[9] Le Chen, Aniko Hannak, Ruijin Ma, and Christo Wilson. Investigating The Impact Of Gender On Rank In Resume Search Engines. In *Annual Conference of the ACM Special Interest Group on Computer Human Interaction*, Montreal, Canada, April 2018.
[10] Consumer Financial Protection Bureau. Using Publicly Available Information To Proxy For Unidentified Race And Ethnicity, 2014. https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf.
[11] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR*, abs/1808.00023, 2018.
[12] Department of Housing and Urban Development. Hud's Implementation Of The Fair Housing Act's Disparate Impact Standard, 2019. https://www.federalregister.gov/documents/2019/08/19/2019-17542/huds-implementation-of-the-fair-housing-acts-disparate-impact-standard.
[13] Emily Dreyfuss. Facebook Changes Its Ad Tech To Stop Discrimination. *WIRED*, 2019. https://www.wired.com/story/facebook-advertising-discrimination-settlement/.
[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012.
[15] Cynthia Dwork and Christina Ilvento. Fairness Under Composition. In *10th Innovations in Theoretical Computer Science Conference (ITCS 2019)*, volume 124 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 33:1–33:20, 2018.
[16] Exhibit A – Programmatic Relief. https://nationalfairhousing.org/wp-content/uploads/2019/03/FINAL-Exhibit-A-3-18.pdf.
[17] Facebook Motion To Dismiss In Onuoha V. Facebook. https://www.courtlistener.com/recap/gov.uscourts.cand.304918/gov.uscourts.cand.304918.34.0.pdf.
[18] Masoomali Fatehkia, Isabelle Tingzon, Ardie Orden, Stephanie Sy, Vedran Sekara, Manuel Garcia-Herranz, and Ingmar Weber. Mapping socioeconomic indicators using social media advertising data. *EPJ Data Science*, 9(1):22, 2020.
[19] Marion Fourcade and Kieran Healy. Classification Situations: Life-chances In The Neoliberal Era. *Accounting, Organizations and Society*, 38(8):559–572, 2013.
[20] David Garcia, Yonas Mitike Kassa, Angel Cuevas, Manuel Cebrian, Esteban Moro, Iyad Rahwan, and Ruben Cuevas. Analyzing gender inequality through large-scale facebook advertising data. *Proceedings of the National Academy of Sciences*, 115(27):6958–6963, 2018.
[21] Rachel Goodman. Why Amazon's Automated Hiring Tool Discriminated Against Women, 2018. https://www.aclu.org/blog/womens-rights/womens-rights-workplace/why-amazons-automated-hiring-tool-discriminated-against.
[22] Jerald Greenberg. A Taxonomy Of Organizational Justice Theories. *Academy of Management review*, 12(1):9–22, 1987.
[23] Aniko Hannak, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. Bias In Online Freelance Marketplaces: Evidence From Taskrabbit And Fiverr. In *ACM Conference on Computer Supported Cooperative Work*, Portland, Oregon, USA, February 2017.
[24] Basileal Imana, Aleksandra Korolova, and John Heidemann. Auditing For Discrimination In Algorithms Delivering Job Ads. In *Proceedings of the Web Conference 2021*, pages 3767–3778, 2021.
[25] Ava Kofman and Ariana Tobin. Facebook Ads Can Still Discriminate Against Women and Older Workers, Despite a Civil Rights Settlement. https://www.propublica.org/article/facebook-ads-can-still-discriminate-against-women-and-older%2Dworkers-despite-a-civil-rights-settlement.
[26] Anja Lambrecht and Catherine Tucker. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management science*, 65(7):2966–2981, 2019.
[27] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting Racial Bias In An Algorithm Used To Manage The Health Of Populations. *Science*, 366(6464):447–453, 2019.
[28] Rebekah Overdorf, Bogdan Kulynych, Ero Balsa, Carmela Troncoso, and Seda Gürses. Questioning The Assumptions Behind Fairness Solutions. *arXiv preprint arXiv:1811.11293*, 2018.
[29] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware Data Mining. In *ACM SIGKDD International Conference of Knowledge Discovery and Data Mining*, Las Vegas, North Dakota, USA, August 2008.
[30] Destiny Peery. The colorblind ideal in a race-conscious reality: The case for a new legal ideal for race relations. *Nw. JL & Soc. Pol'y*, 6:473, 2011.
[31] Piotr Sapiezyński, Valentin Kassarnig, Christo Wilson, Sune Lehmann, and Alan Mislove. Academic Performance Prediction In A Gender-imbalanced Environment. In *Workshop on Responsible Recommendation*, Como, Italy, August 2017.
[32] Traci Schlesinger. The failure of race neutral policies: How mandatory terms and sentencing enhancements contribute to mass racialized incarceration. *Crime & delinquency*, 57(1):56–81, 2011.
[33] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness And Abstraction In Sociotechnical Systems. In *Conference on Fairness, Accountability, and Transparency*, Atlanta, Georgia, USA, January 2019.
[34] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. On The Potential For Discrimination In Online Targeted Advertising. In *Conference on Fairness, Accountability, and Transparency*, New York, New York, USA, February 2018.
[35] Upturn Amicus Brief In Onuoha V. Facebook. https://www.courtlistener.com/recap/gov.uscourts.cand.304918/gov.uscourts.cand.304918.76.1.pdf.
[36] Samuel Yeom, Anupam Datta, and Matt Fredrikson. Hunting For Discriminatory Proxies In Linear Regression Models. In *Advances in Neural Information Processing Systems*, pages 4568–4578, 2018.