Algorithms That "Don't See Color": Comparing Biases in Lookalike and Special Ad Audiences

Piotr Sapiezynski
Northeastern University
p.sapiezynski@northeastern.edu

Avijit Ghosh
Northeastern University
avijit@ccs.neu.edu

Levi Kaplan
Northeastern University
kaplan.l@husky.neu.edu

Alan Mislove
Northeastern University
amislove@ccs.neu.edu

Aaron Rieke
Upturn
aaron@upturn.org

ABSTRACT

Today, algorithmic models are shaping important decisions in domains such as credit, employment, healthcare, and criminal justice. At the same time, researchers and journalists have repeatedly shown that these algorithms can have discriminatory effects. Some organizations have tried to mitigate these effects by simply removing demographic features from an algorithm's inputs. If an algorithm is not provided with a demographic feature, one might think, then its outputs should not discriminate with respect to that feature. This may not be true, however, when there are other features that are *correlated* with that demographic feature. Today, there are few public experiments that measure how removing demographic inputs affects the outputs of large-scale, real-world algorithmic systems.

In this paper, we explore the limits of this approach using a unique opportunity created by a recent lawsuit settlement concerning discrimination on Facebook's advertising platform. In 2019, Facebook agreed to modify its *Lookalike Audiences* tool—which creates target sets of users (audiences) for ads by identifying users who share "common qualities" with users in a source audience provided by an advertiser—by removing certain demographic features as inputs to its algorithm. The modified tool, called *Special Ad Audiences*, is presumably intended to reduce the potential for discrimination in target audiences. We create a series of Lookalike and Special Ad audiences based on *biased* source audiences—i.e., source audiences that have known skew along the lines of gender, age, race, and political leanings—and show that the resulting Lookalike and Special Ad audiences both reflect these biases. Importantly, these biases are present *despite* the fact that Special Ad Audiences algorithm is not provided with the demographic features along which our source audiences are skewed.

Our results suggest that, relative to Lookalike Audiences, Special Ad Audiences do little to reduce demographic biases in target audiences. More broadly, we provide experimental proof that merely removing demographic features from a real-world algorithmic system's inputs can fail to prevent biased outputs. Organizations using algorithms to help mediate access to important life opportunities should consider other approaches to mitigating discriminatory effects.

1 INTRODUCTION

Organizations are now using algorithmic models[1] ("algorithms") in a variety of important domains, including healthcare [36], credit [21], employment [10, 27], and content distribution [4]. Sometimes, these algorithms can be beneficial. But too often, they can lead to discriminatory effects. For example, in the context of criminal justice, ProPublica showed that the COMPAS risk-assessment tool [9] used by judges to help make bail decisions was particularly likely to mislabel Black defendants as future criminals [12]. Similarly, facial recognition algorithms have been shown to perform significantly worse for Black women [8]. Facebook, in its quest to show relevant ads to users, diverts the delivery of employment and housing ads away from some demographic groups, even when an advertiser is trying to reach a broad, diverse audience [4]. These kinds of discriminatory effects can be challenging to detect, measure, and articulate.

Some have proposed mitigating discriminatory effects by removing demographic features from an algorithm's inputs. For example, as we discuss in more detail in Section 5.1, the U.S. Department of Housing and Urban Development (HUD) recently proposed a rule that would apply this approach to housing discrimination [24]. This approach is flawed, however, because algorithms can effectively use omitted demographic features by combining other inputs that are each *correlated* with the those features, potentially nullifying any protection from discriminatory effects. This is particularly true in large-scale machine learning (ML) systems, which take can as input hundreds or thousands 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* [19] 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 [15] Facebook over violations of the Fair Housing Act (FHA) [1]. 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 *Special Ad Audiences* [3], which works like Lookalike Audiences, except its

[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.

algorithm does *not* consider user profile fields of "age, gender, relationship status, religious views, school, political views, interested in, or zip code" when detecting common qualities [28].

We seek to learn whether the Special Ad Audience algorithm actually produces significantly less *biased* audiences than the Lookalike Audience algorithm. 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 in the settlement, enabling us to examine whether simply removing the protected features as input to an algorithm is sufficient eliminate bias 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, race, and political views. Our results show that:

- For gender, our Special Ad audiences[2] are biased 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 the same ad targeted using Special Ad audiences delivered to 91.2% women.
- For age, our Special Ad audiences are almost as biased as Lookalike audiences when using source audiences that are from a single age range or a controlled mix of two age ranges.
- For race, we use a different methodology to estimate the racial makeup of the Lookalike and Special Ad Audiences, as Facebook does not report the delivery along racial lines. Our results suggest that Special Ad Audiences can skew along racial lines, as is true for Lookalike Audiences.
- For political views, we use a similar methodology and find a similar skew for both Lookalike and Special Ad Audiences. However, we observed less overall bias along political views than race in both audience types.
- 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 Special Ad Audiences. Using a source audience consisting of Facebook employees—identified by `@fb.com` email addresses—we find that the resulting Special Ad audience skews heavily towards 25–34-year-old men.
- We confirm that previous findings on how Facebook's delivery mechanisms can cause further skews in who is shown ads hold for Special Ad Audiences. We show that an ad for artificial intelligence jobs delivers mostly to young men, while an ad for supermarket jobs delivers mostly to middle-aged women, despite targeting the same gender- and age-balanced Special Ad audience.

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 with respect to those features. This work also demonstrates a methodology by which other algorithms could be studied.

[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 created using the tool.

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.

The remainder of this paper is organized as follows: Section 2 provides background on Facebook's ad targeting tools and related work. Section 3 introduces our methodology and Section 4 presents our results. Section 5 provides a concluding discussion.

Ethics We took careful consideration of ethics when conducting the research in this paper. *First*, we minimized harm to Facebook users by only running "real" ads, i.e., if a user happened to click on one of our ads, they were sent by Facebook to a real-world site relevant to content the ad. We therefore did not have any direct interaction with the users who were shown our ad, and did not collect any of their personally identifying information. *Second*, we minimized harm to Facebook by running and paying for our ads just like any other advertiser. In cases where we were running employment ads, we flagged them as such using Facebook's tools.

2 BACKGROUND

In this section, we provide background on Facebook's ad targeting tools, including Special Ad Audiences, and 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* [20], including over 1,000 demographic, behavioral, and interest-based features.

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* [47], 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 [13]. 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* [19], which creates an audience of users who share "common qualities" with users in a Custom audience provided by the advertiser (called the *source audience*). A screenshot of Facebook's advertiser interface is shown in Figure 1 (left); the advertiser must select the country where they wish Facebook to select users from ("Audience Location") and then must select the fraction of that country's population to include in the new Lookalike Audience (ranging from 1% to 10%). Our prior work has demonstrated that Lookalike Audiences can reproduce demographic skews present in source audiences [43].

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 [15]. The lawsuit was settled in March 2019, and Facebook agreed to make a number of changes

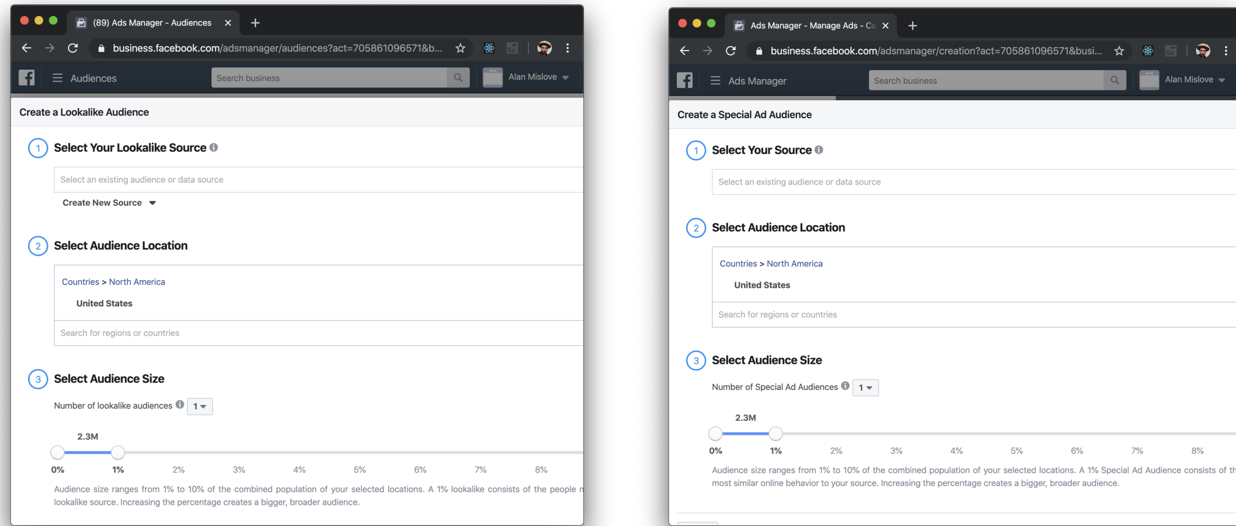

Figure 1: Screenshots of creation process for both Lookalike Audiences (left) and Special Ad Audiences (right). Both of them are the same from the advertiser's perspective: The advertiser first selects a source Custom audience, then selects a target country, and finally selects the fraction of that country's users to include in the new audience.

to its ad targeting tools. Relevant here, Facebook agreed to change how Lookalike Audiences (LAL) works when used with housing, credit, and employment (HEC) ads [17]:

> 5. Lookalike Audience ("LAL"): In the HEC Flow, LAL tool and marketing will be modified as follows:
>
> (a) LAL tool may consider the following user profile fields: country, region, profession and field of study. LAL tool will not consider the following user profile fields: age, gender, relationship status, religious views, school, political views, interested in, or zip code.

Facebook now refers to this modified Lookalike Audiences tool as *Special Ad Audiences* [3]. Facebook says [28]:

> [Special Ad Audiences] will create an audience based on similarities in online behavior and activity but that does not use certain categories, including age, gender, ZIP code or other similar categories.

From an advertiser's perspective, Special Ad Audiences are created just like Lookalike Audiences (i.e., based on a source Custom audience). A screenshot of Facebook's interface for creating Lookalike Audiences is shown in Figure 1 (left), and for Special Ad Audiences in Figure 1 (right).

2.3 Related work

We briefly overview related work on studying and mitigating algorithmic bias, as well as Facebook's advertising platform.

Algorithmic bias Concerns over bias in algorithms have galvanized a growing research community. This community has developed a number of approaches to *algorithmic auditing* [40], a process of seeking to understand an algorithm's inputs, outputs, and potential for discriminatory effects. Researchers have successfully studied a variety of widely deployed algorithmic systems including face-recognition systems [8], e-commerce sites [26], search engines [14, 25, 31, 32, 38], job seeking sites [10, 27], online translation services [7], or health-management [36]. A number of proposals have been put forward to mitigate the potential algorithmic biases; we refer the reader to a survey of both sources of bias and mitigation approaches [35] for a more in-depth treatment.

We highlight a few works most closely related to our topic of measurement. Greenberg distinguishes two kinds of fairness concerns, *distributive* and *procedural* [30]. The former aims to assure balanced outcomes, whereas the latter focuses on the process itself. Elimination of features from an algorithm's input (as with Special Ad Audiences) falls into the *procedural* category. Grgić-Hlača et al. [23] propose a framework which relies on human moral judgments to determine which features are fair to use. They point out that while people can accurately judge relevance and privacy aspects of a feature in decision making, they tend to fail at predicting the impact that feature might have on the decision outcomes. Specifically, certain features might appear fair to human judges even though they are correlated with sensitive features. In such cases process fairness does not lead to outcome fairness, and additional constraints must be enforced. Further, there are cases in which none of the features is a strong proxy for a sensitive attribute but features can form a proxy when combined [11, 39]. Finally, even if none of the features or their combinations are unfair, their predictive performance might be different across sub-populations. Then, in an effort to to minimize the total error, the classifier will fit the majority group better than the minority [29, 41]. Taken together, these prior works paint a clear picture of process fairness as insufficient to ensure fair outcomes.

More recently, there has been a growing agreement among scholars that focusing on particular algorithms is too narrow of a problem definition. Real-world algorithmic decision systems are often composed of multiple algorithmic subsystems and can be discriminatory as a whole, even if built from a series of fair algorithms [16] Algorithms need to be modeled along with the other components of the *sociotechnical* systems they are embedded in [42]. The burden of these investigations lies on independent researchers and auditors since the companies who operate these algorhtms might not be incentivized to measure and address the externalities they cause [37].

Facebook's advertising platform Facebook runs one of the world's most powerful advertising platforms, and has been the object of study for a number of research projects. Prior work has demonstrated that Facebook was using PII provided for security features (e.g., two-factor authentication) was used to allow advertisers to target users with ads [46], that Custom Audiences can be used to leak user's PII [45], that Facebook's ad targeting options offer a variety of mechanisms to create discriminatory audiences [43], that political advertisers on Facebook with higher budgets target people using more privacy sensitive features [22] and that Facebook's ad delivery system *itself* may introduce unwanted biases when deciding which users should be presented with life opportunity [4, 33] and political ads [5].

3 METHODOLOGY

We now describe the methodology we use to study Lookalike and Special Ad Audiences. Recall that our goal is to measure whether Special Ad Audiences produce significantly less biased audiences than Lookalike Audiences. We therefore need to be able to generate source audiences with controlled and known skew, from which we can create a Lookalike and a Special Ad audience. To do so, we re-use an approach from prior work [4], relying on voter records from New York and North Carolina. These records are available to the public, and include 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. For both types, we select users in the U.S. and choose the smallest size option (1% of the population).

For some of our experiments, to measure the makeup of a target audience, we run actual ads and 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). Unless otherwise noted, we create a generic ad for Google Web Search, which has basic text ("Search the web for information") and a link to Google Search. 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. Thus, we are able to run the same, generic ad creative using both Lookalike and Special Ad audiences.

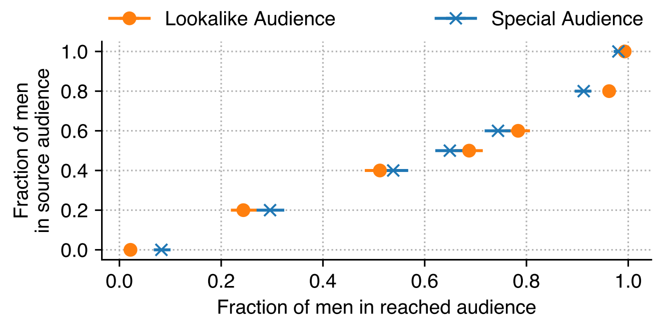

Figure 2: Gender breakdown of ad delivery to Lookalike and Special Ad audiences created from the same source audience with varying fraction of male users, using the same ad creative. We can observe that both Lookalike and Special Ad audiences reflect the gender distribution of the source audience, despite the lack of gender being provided as an input to Special Ad Audiences.

4 RESULTS

We now present our experiments and analyze the results. We first examine whether Lookalike and Special Ad Audiences can be biased along the lines of gender and age, which are straightforward to measure as Facebook provides delivery statistics in the advertiser interface. Next, we focus on race and political views, using a different methodology. Finally, we show the real-world implications of these experiments using a series of employment and credit ads.

4.1 Gender

We begin by focusing on gender. We create seven Custom audiences based on New York voter records. Each audience contains 10,000 individuals, with varying fractions of men: 0%, 20%, 30%, 40%, 50%, 60%, 80%, 100%. We then 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 2 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 skew has been observed by prior work [4, 33] 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, we observe that a similar level of bias (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. Overall,

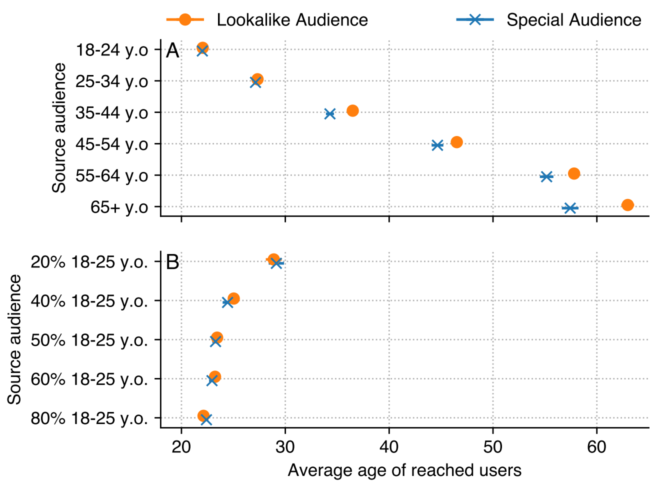

Figure 3: Age breakdown of ad delivery to Lookalike and Special Ad audiences created from the same source audience, using the same ad creative. We can observe extremely similar levels of bias, despite the lack of age as an input to Special Ad audiences. Panel A shows the results for source audiences consisting only of users in one age bracket. Panel B shows the results of mixing the youngest and the oldest users in different proportions.

the Special Ad audiences show a bit less bias when compared to the Lookalike audiences, but the trend is clear.

4.2 Age

Next, we turn to age. Facebook reports delivery based on users' age in terms of fixed ranges: 18–24, 25–34, 35–44, 45–54, 55–64, 65+. We therefore design an experiment based off of these ranges. We create biased Custom audiences by selecting voter records that consist *only* of 10,000 users within a single age range. As before, we then run two ads for each Custom audience: one to a Lookalike audience and another to a Special Ad audience.

Figure 3A presents the results, and the top six rows of Table A1 provide a more detailed breakdown (and includes the average age of the delivery audience in the final column). *First*, we can immediately observe that that average age[3] for both the Lookalike audience and the Special Ad audience increases in the same manner, with the average age being 21.9 for the audiences created from the 18–24 age range and between 57 and 63 for the audiences created from the 65+ age range. *Second*, examining the distribution of delivery to each age range shows very similar trends in each pair of audiences, with the 18–24 source audience delivering almost exclusively to young Facebook users in both cases, and the 65+ source audience delivering primarily to older Facebook users in both cases as well. *Third*, as with gender, we may be observing that Special Ad audiences are slightly less biased—the average age of Special Ad audiences is closer to the median Facebook user age in all of the older user groups compared to that of the Lookalike audiences—but the overall effect is very strong.

Additionally, we look at what happens if we "mix" together users from different age ranges, as opposed to running experiments with a source audience from a single age range. Specifically, we mix together both old (65+) and young (18–24) users in different proportions and examine the ultimate delivery audience for the Lookalike and Special Ad audiences. Figure 3B and the bottom five rows of Table A1 present these results in the same format before. As before, we can see that the average age follows a very similar trend for both types of audiences, and that the breakdown in delivery is quite similar as well. Thus, the effect we are observing is not limited to homogeneous source audiences.

4.3 Race

Next, we turn to examine the extent to which Special Ad Audiences can be biased along racial lines, in the same manner we have observed Lookalike Audiences to be in past work [43]. 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* [48], which is an estimate provided by Facebook of the number of users that match the advertiser's targeting criteria and can be reached within the specified budget. We set the daily budget to the maximum allowed value ($1M) to best approximate the 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.[4]

The procedure has two steps: audience creation and targeting.

Audience Creation. As before, we start with voter records from North Carolina (which provide race information). We focus on two racial groups: Black (defined as users who self-report as Non-Hispanic Black) and white (defined as users who self-report as Non-Hispanic white). For each race, 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. We refer to these audiences as `w_10k` and `w_900k` (for the white audiences) and `b_10k` and `b_900k` (for the Black audiences).

Next, we use the Custom audiences with 10,000 users to create corresponding Lookalike and Special Ad audiences. We refer to these 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 $S_{\texttt{b\_10k}}$. Our goal is then to estimate the racial bias of these Lookalike and Special Ad audiences.

Targeting. We now use the ad targeting interface to obtain such estimates. To do so, we begin the ad creation process and set our budget is the maximal value allowed ($1M/day). We also specify that we only target users in North Carolina.

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 large 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,

[3] We estimated the average age using weighted midpoints of each age range; it should be viewed simply as a summary of the aggregate distribution and not the precise average age.

[4] We repeated the procedure using both the midpoint as well as the upper value and found similar results.

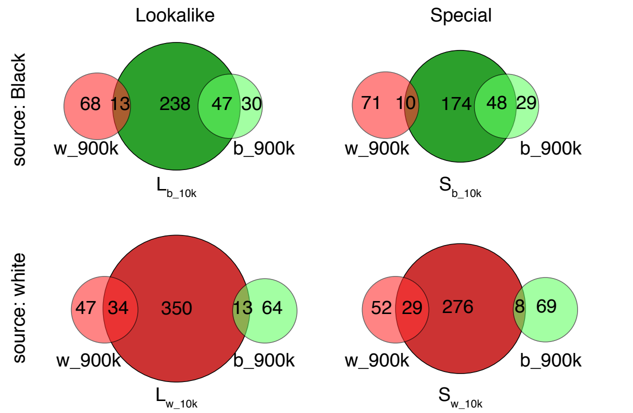

Figure 4: Venn diagrams of overlap between Lookalike and Special audiences created from samples of 10,000 Black and white voters. While we do not know the race of the vast majority of the created audiences, the part that we do know shows racial bias.

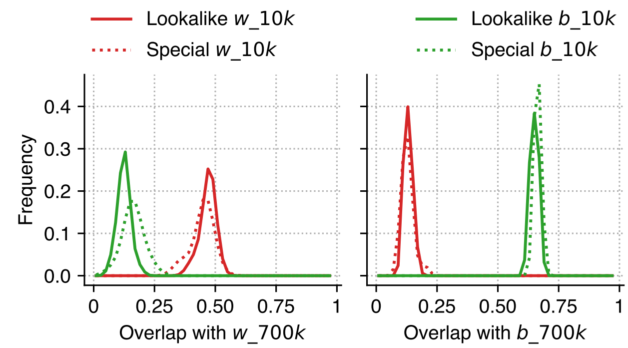

Figure 5: Both Lookalike and Special Ad audiences created from source audiences of white users containing a higher fraction of white users than Black users. Conversely, audiences created from source audiences of Black users contain a higher fraction of Black users than white users.

and record the potential daily reach (e.g., 360,000). Now, we can observe that excluding w_900k from $L_{\mathtt{w\_10k}}$ caused the potential daily reach to drop by 37,000, indicating that approximately 46% (37,000/81,000) of w_900k were present in $L_{\mathtt{w\_10k}}$. We can then repeat the process with excluding b_900k, and measure the fraction of the large Black audience that is present in $L_{\mathtt{w\_10k}}$. By comparing the fraction of w_900k and b_900k that are present in $L_{\mathtt{w\_10k}}$, we obtain an estimate of the racial bias of $L_{\mathtt{w\_10k}}$.

Limitations. It is important to note that, unlike in our experiments with gender and age, here we do not know the race of a vast majority of the audience. This is because the Lookalike and Special Ad audiences that Facebook creates consist mostly of people who appear not to be in our voter records. 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, as further emphasized in Figure 4. However, these estimates do give us a small window into the makeup of these Lookalike and Special Ad audiences.

Results. We begin by presenting Venn diagrams in Figure 4 that capture the overlap between all of the audiences. We summarize

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

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 b_900k but only 16% of w_900k. 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 North Carolina). Thus, our results are not conclusive—but only suggestive—that the overall audiences are similarly biased.

Robustness. Here, we confirm that the presented results are robust to the random selection of seed from which Lookalike and Special Ad audiences are created. To this end, we repeat the described process with 20 non-overlapping b_10k and 20 non-overlapping w_10k audiences. We create a Lookalike and a Special Ad audience for each and compute the overlap with large b_700k and b_700k audiences and report the resulting fractions in Figure 5. We note that the racial skew observable in Lookalike audiences persists in Special Ad audiences but the effect is slightly smaller.

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.

Specifically, we create source audiences of Republicans and Democrats (r_10k and d_10k), as well as large Republican and Democrat audiences (r_900k and d_900k). We then use the source

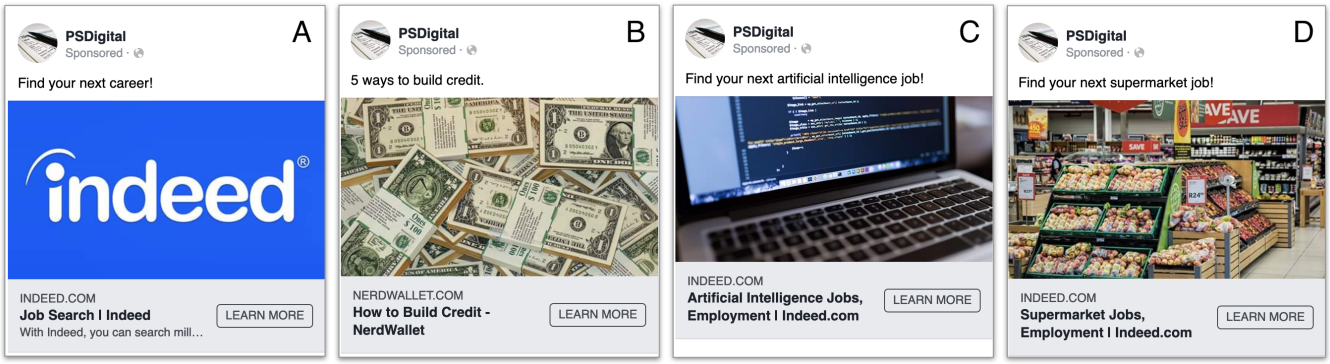

Figure 6: Ad creatives used throughout the paper. All of our ads linked directly to the domains shown in the ad.

audiences to create both Lookalike audiences ($L_{\texttt{r\_10k}}$ and $L_{\texttt{d\_10k}}$) and Special Ad audiences ($S_{\texttt{r\_10k}}$ and $S_{\texttt{d\_10k}}$). As with race, we run the same generic ad to all audiences, and examine the fraction of the large audiences that are present in the Lookalike and Special Ad audiences.

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.

Source	Type	Percent overlap Democrat (`d_900k`)	Republican (`r_900k`)
100% Democrat	Lookalike $L_{\texttt{d\_10k}}$	51.6	31.8
	Special $S_{\texttt{d\_10k}}$	42.2	25.8
100% Republican	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.

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 without dramatic 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 source audience, 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:

(1) We create a Custom audience consisting of randomly generated American phone numbers, 11,000 of which Facebook matched to existing users. This is our baseline audience that we use to measure the bias in the created audience.
(2) We create another Custom audience consisting of 12,356,604 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.
(3) We create Special Ad audiences based on each of these two Custom audiences.
(4) We run two generic job search ads (see Figure 6A), each to one of these Special Ad audiences, at the same time, from the same account, with the same budget. This way we eliminate the side effects of optimization based on the ad content or budget.
(5) We collect the delivery statistics with age/gender breakdown.

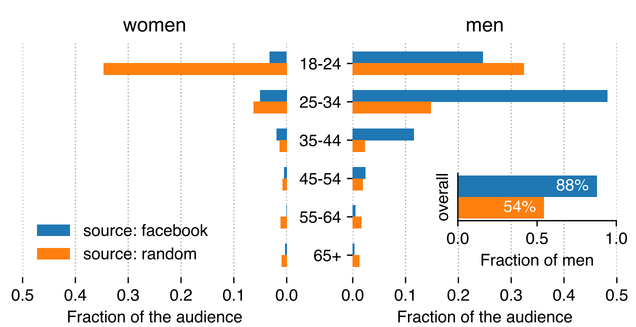

Figure 7: Gender and age breakdown of a generic job ad delivery to a Special Ad audience based on random American users (in orange) and a Special Ad audience based on Facebook employees (in blue). The based on Facebook employees is predominantly male and 25-34.

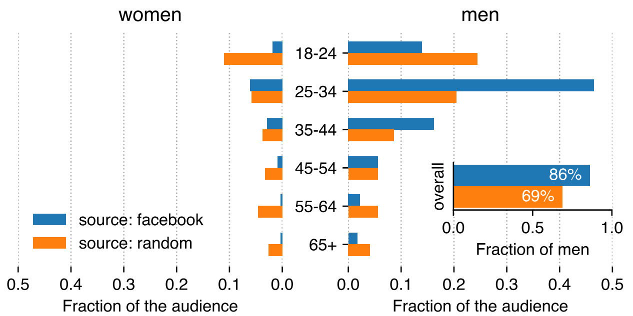

Figure 8: Gender and age breakdown of delivery of credit building ads to a Special Ad audience based on random American users (in orange) and a Special Ad audience based on Facebook employees (in blue). The Facebook-based audience still delivers predominantly males aged 25-34, but the random audience skews more male too.

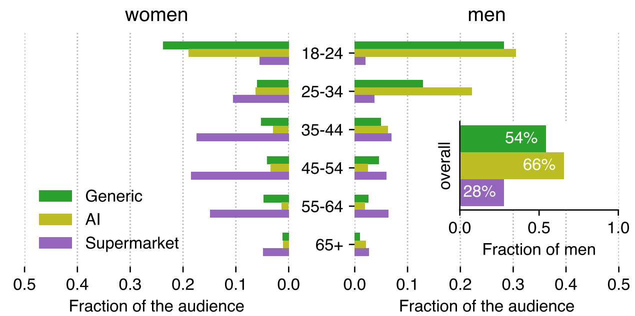

Figure 9: Gender and age breakdown of delivery of job ads to a Special Ad audience based on random American users. We observe that even if the source audience is held constant and is approximately gender-balanced, the content of the ad can still lead to large skews: ads for supermarket jobs deliver to 72% female audience, whereas ads for jobs in AI delivery to 66% male audience. Note also the skews in age: supermarket jobs deliver mostly to people aged 35 and older, whereas the AI jobs deliver nearly exclusively to people younger than 35.

Figure 7 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. Finally, 47% of all deliveries to the Facebook Special Ad audience are to users in California, compared to 2% in the baseline audience.[5] 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).

We confirm these findings by running an additional experiment in the "credit" category, advertising tips for building credit (see ad copy in Figure 6B) to the same two Special Ad audiences. The results of this experiment are presented in a similar format in Figure 8. We observe that the ad targeting the Facebook-based audience still delivers predominantly towards male and 25–34-year-old users. However, we also observe a shift towards delivering to male users in the random audience, even though it is the same audience as in the generic job ad. This is likely an effect of the ad creative on how Facebook algorithms estimate relevance to different groups of people [4]. We further explore these effects in the next section.

4.6 Content-based skew in delivery

In previous work [4], we 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 we held the target audience 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 re-use the Special Ad audience created from the random 11,000 users which we expect to be approximately gender- and age-balanced. We then run generic job ads along with ads for supermarket and artificial intelligence pointing to search for either keyword on `indeed.com`, see Figure 6A, C, and D. We report the fraction of men in the reached audience in Figure 9; we can immediately observe 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). This skew when delivering ads to a gender-balanced audience underlines 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 DISCUSSION

We have demonstrated that both Lookalike and Special Ad Audiences can create similarly biased target audiences from the same source audiences. To reiterate, we are not claiming that Facebook has incorrectly implemented Special Ad Audiences, nor are we suggesting that the company has violated its settlement agreement. Rather, our findings are a result 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.

---

[5] While matching based on state is not prohibited in the settlement, these numbers indicate that our method of selecting Facebook employees based on random email addresses `@fb.com` is correct.

5.1 Policy implications

In the U.S., President Trump's administration has has directed a range of civil rights officials to reexamine how 'disparate impact' regulations might be changed or removed [34]. Disparate impact is a legal doctrine that says facially neutral practices in employment, housing, credit, and other areas can still lead to a finding of discrimination if they adversely affect a protected group. To date, at least one regulator is considering new rules that would insulate companies from disparate impact liability if their algorithms do not rely on protected demographic factors. Our results strongly indicate that this is a flawed approach.

In August 2019, HUD published a proposed rule [24] to amend its interpretation of the FHA in light of the U.S. Supreme Court's ruling in *Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc.* In particular, the proposed rule states in §100.500(c)(2)(i) and (iii) that a defendant can claim a plaintiff has failed to establish a prima facie case for disparate impact when:

> ... a plaintiff alleges that the cause of a discriminatory effect is a model used by the defendant, such as a risk assessment algorithm, and the defendant:
> (i) Provides the material factors that make up the inputs used in the challenged model and shows that these factors do not rely in any material part on factors that are substitutes or close proxies for protected classes under the Fair Housing Act and that the model is predictive of credit risk or other similar valid objective;

This describes, in essence, the approach that the settlement proscribed and led to the creation of Special Ad Audiences. Our results reinforce what dozens of companies, advocates, and researchers told HUD when its proposed rule was open for public comment: Removing protected features from an algorithm's inputs is not enough to prevent discriminatory effects.

5.2 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 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) [2] 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" [18], 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 [44].

ACKNOWLEDGEMENTS

The authors thank Ava Kofman and Ariana Tobin for suggesting the experiments presented in Sections 4.5 & 4.6 as well as for going an extra mile (or two) for their ProPublica story around this work. 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-1616234 and CNS-1916153, and Mozilla Research Grant 2019H1.

REFERENCES

[1] 24 CFR § 100.75 – Discriminatory advertisements, statements and notices. https://www.law.cornell.edu/cfr/text/24/100.75.

[2] 47 USC § 230 – Protection for private blocking and screening of offensive material. https://www.law.cornell.edu/uscode/text/47/230.

[3] About Audiences for Special Ad Categories. https://www.facebook.com/business/help/2220749868045706.

[4] Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., and Rieke, A. Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proceedings of the ACM on Human-Computer Interaction 3*, CSCW (2019), ACM.

[5] Ali, M., Sapiezynski, P., Korolova, A., Mislove, A., and Rieke, A. Ad Delivery Algorithms: The Hidden Arbiters of Political Messaging. *arXiv preprint arXiv:1912.04255* (2019).

[6] Blass, J. Algorithmic Advertising Discrimination. *Northwestern University Law Review 114*, 2 (2019).

[7] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Neural and Information Processing Systems* (Barcelona, Spain, Dec. 2016).

[8] Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability, and Transparency* (New York, New York, USA, Feb. 2018).

[9] COMPAS Classification Tool. https://www.equivant.com/northpointe-suite/.

[10] Chen, L., Hannak, A., Ma, R., and Wilson, C. 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).

[11] 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.

[12] Courtland, R. Bias detectives: the researchers striving to make algorithms fair. *Nature 558*, 7710 (2018).

[13] Custom Audiences from your Customer List. https://www.facebook.com/business/help/606443329504150.

[14] Diakopoulos, N., Trielli, D., JenniferStark, and Mussenden, S. I Vote For—How Search Informs Our Choice of Candidate. In *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple* (2018), Oxford University Press, M. Moore and D. Tambini, eds.

[15] Dreyfuss, E. Facebook Changes Its Ad Tech to Stop Discrimination. In *WIRED*. https://www.wired.com/story/facebook-advertising-discrimination-settlement/.

[16] Dwork, C. and Ilvento, C. Fairness Under Composition. In *10th Innovations in Theoretical Computer Science Conference (ITCS 2019)*, vol. 124 of *Leibniz International Proceedings in Informatics (LIPIcs)* (2018).

[17] Exhibit a – Programmatic Relief. https://nationalfairhousing.org/wp-content/uploads/2019/03/FINAL-Exhibit-A-3-18.pdf.

[18] Facebook Motion to Dismiss in Onuoha v. Facebook. https://www.courtlistener.com/recap/gov.uscourts.cand.304918/gov.uscourts.cand.304918.34.0.pdf.

[19] Facebook: About Lookalike Audiences. https://www.facebook.com/business/help/164749007013531.

[20] Facebook: Core Targeting. https://www.facebook.com/business/help/633474486707199.

[21] Fourcade, M. and Healy, K. Classification situations: Life-chances in the neoliberal era. *Accounting, Organizations and Society 38*, 8 (2013).

[22] Ghosh, A., Venkatadri, G., and Mislove, A. Analyzing Facebook Political Advertisers' Targeting. In *Workshop on Technology and Consumer Protection* (San Francisco, California, USA, May 2019).

[23] Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., and Weller, A. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *AAAI Conference on Artificial Intelligence* (New Orleans, Louisiana, USA, Feb. 2018).

[24] HUD's Implementation of the Fair Housing Act's Disparate Impact Standard. https://www.federalregister.gov/documents/2019/08/19/2019-17542/huds-implementation-of-the-fair-housing-acts-disparate-impact-standard.

[25] Hannak, A., Sapiezyński, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D., Mislove, A., and Wilson, C. Measuring Personalization of Web Search. In

Source	Type	Delivery audience age breakdown (%) 18–24	25–34	35–44	45–54	55–64	65+	Average Age
100% 18–24	Lookalike	88.7	10.9	0.2	0.0	0.0	0.2	22.0
	Special	88.3	11.5	0.2	0.0	0.0	0.0	22.0
100% 25–34	Lookalike	30.0	66.9	2.9	0.1	0.1	0.1	27.0
	Special	33.0	62.9	3.9	0.1	0.1	0.0	26.8
100% 35–44	Lookalike	3.2	34.0	54.8	7.0	0.6	0.5	36.0
	Special	6.8	49.4	36.1	6.2	1.0	0.6	33.8
100% 45–54	Lookalike	0.7	5.1	32.7	49.7	9.0	3.0	46.1
	Special	3.9	14.1	32.4	32.6	11.3	5.8	44.3
100% 55–64	Lookalike	1.2	2.5	7.3	20.6	39.6	28.9	57.8
	Special	2.4	6.2	11.7	21.0	31.3	27.5	55.1
100% 65+	Lookalike	3.3	4.2	3.2	3.3	19.6	66.4	63.5
	Special	6.7	9.6	5.2	6.9	23.2	48.5	57.7
20% 18–24, 80% 65+	Lookalike	70.1	14.5	1.1	0.9	3.3	10.0	28.9
	Special	65.1	18.1	2.3	2.4	4.4	7.7	29.1
40% 18–24, 60% 65+	Lookalike	78.6	14.8	0.7	0.5	1.5	3.9	25.0
	Special	80.3	13.9	1.2	0.5	1.2	2.9	24.4
50% 18–24, 50% 65+	Lookalike	84.1	12.5	0.6	0.1	0.6	2.0	23.4
	Special	81.3	16.0	0.7	0.6	0.5	0.9	23.2
60% 18–24, 40% 65+	Lookalike	83.2	14.1	0.4	0.1	0.9	1.2	23.2
	Special	83.3	14.9	0.8	0.4	0.3	0.7	22.9
80% 18–24, 20% 65+	Lookalike	89.1	10.1	0.4	0.1	0.1	0.3	22.1
	Special	85.3	13.8	0.6	0.1	0.2	0.0	22.3

Table A1: Breakdown in age of delivery audience of ads to Lookalike and Special Ad Audiences created from the same source audience, using the same ad creative. The top six rows represent source audiences with a single age group; the bottom five rows represent source audiences with a mix of young and old users.

International World Wide Web Conference (Rio de Janeiro, Brazil, May 2013).

[26] Hannak, A., Soeller, G., Lazer, D., Mislove, A., and Wilson, C. Measuring Price Discrimination and Steering on E-commerce Web Sites. In *ACM Internet Measurement Conference* (Vancouver, Canada, Nov. 2014).

[27] Hannak, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., and Wilson, C. Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr. In *ACM Conference on Computer Supported Cooperative Work* (Portland, Oregon, USA, Feb. 2017).

[28] Help: Choosing a Special Ad Category. https://www.facebook.com/business/help/298000447747885.

[29] Irene, C., D, J. F., and David, S. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems* (2018).

[30] Jerald, G. A taxonomy of organizational justice theories. *Academy of Management review 12*, 1 (1987), Academy of Management Briarcliff Manor, NY 10510.

[31] Kay, M., Matuszek, C., and Munson, S. A. Unequal representation and gender stereotypes in image search results for occupations. In *Annual Conference of the ACM Special Interest Group on Computer Human Interaction* (2015).

[32] Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K., and Karahalios, K. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. In *ACM Conference on Computer Supported Cooperative Work* (Portland, Oregon, USA, Feb. 2017).

[33] Lambrecht, A. and Tucker, C. E. Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads. (2018). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2852260.

[34] Meckler, L. and Barrett, D. Trump administration considers rollback of anti-discrimination rules. In *Washington Post* (2019). https://www.washingtonpost.com/local/education/trump-administration-considers-rollback-of-anti-discrimination-rules/2019/01/02/f96347ea-046d-11e9-b5df-5d3874f1ac36_story.html.

[35] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. https://arxiv.org/abs/1908.09635.

[36] Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366*, 6464 (2019).

[37] Overdorf, R., Kulynych, B., Balsa, E., Troncoso, C., and Gürses, S. Questioning the assumptions behind fairness solutions. *arXiv preprint arXiv:1811.11293* (2018).

[38] Robertson, R. E., Jiang, S., Joseph, K., Friedland, L., Lazer, D., and Wilson, C. Auditing Partisan Audience Bias within Google Search. In *Annual Conference of the ACM Special Interest Group on Computer Human Interaction* (2018).

[39] Samuel, Y., Anupam, D., and Matt, F. Hunting for discriminatory proxies in linear regression models. In *Advances in Neural Information Processing Systems* (2018).

[40] Sandvig, C., Hamilton, K., Karahalios, K., and Langbort, C. Auditing algorithms: Research methods for detecting discrimination on internet platforms. In *International Communication Association Conference* (2014).

[41] Sapiezyński, P., Kassarnig, V., Wilson, C., Lehmann, S., and Mislove, A. Academic performance prediction in a gender-imbalanced environment. In *Workshop on Responsible Recommendation* (Como, Italy, Aug. 2017).

[42] Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. Fairness and abstraction in sociotechnical systems. In *Conference on Fairness, Accountability, and Transparency* (Atlanta, Georgia, USA, Jan. 2019).

[43] Speicher, T., Ali, M., Venkatadri, G., Ribeiro, F. N., Arvanitakis, G., Benevenuto, F., Gummadi, K. P., Loiseau, P., and Mislove, A. On the Potential for Discrimination in Online Targeted Advertising. In *Conference on Fairness, Accountability, and Transparency* (New York, New York, USA, Feb. 2018).

[44] Upturn Amicus Brief in Onuoha v. Facebook. https://www.courtlistener.com/recap/gov.uscourts.cand.304918/gov.uscourts.cand.304918.76.1.pdf.

[45] Venkatadri, G., Liu, Y., Andreou, A., Goga, O., Loiseau, P., Mislove, A., and Gummadi, K. P. Privacy Risks with Facebook's PII-based Targeting: Auditing a Data Broker's Advertising Interface. In *IEEE Symposium on Security and Privacy* (San Francisco, California, USA, May 2018).

[46] Venkatadri, G., Lucherini, E., Sapiezyński, P., and Mislove, A. Investigating sources of PII used in Facebook's targeted advertising. In *Privacy Enhancing Technologies Symposium* (Stockholm, Sweden, July 2019).

[47] What's a Custom Audience from a Customer List? https://www.facebook.com/business/help/341425252616329/.

[48] What's the Difference Between Estimated Daily Reach and Potential Reach? https://www.facebook.com/business/help/1438142206453359.

APPENDIX

We present detailed results for the age experiment in Table A1.