Measurement and Analysis of Implied Identity in Ad Delivery Optimization

Levi Kaplan
Northeastern University
Boston, MA, USA

Nicole Gerzon
Northeastern University
Boston, MA, USA

Alan Mislove
Northeastern University
Boston, MA, USA

Piotr Sapiezynski
Northeastern University
Boston, MA, USA

ABSTRACT

Online services such as Facebook and Google serve as a popular way by which users today are exposed to products, services, viewpoints, and opportunities. These services implement advertising platforms that enable precise *targeting* of platform users, and they optimize the *delivery* of ads to the subset of the targeted users predicted to be most receptive. Unfortunately, recent work has shown that such delivery can—often without the advertisers' knowledge—show ads to biased sets of users based only on the content of the ad. Such concerns are particularly acute for ads that contain pictures of people (e.g., job ads showing workers), as advertisers often select images to carefully convey their goals and values (e.g., to promote diversity in hiring). However, it remains unknown how ad delivery algorithms react to—and make delivery decisions based on—demographic features of people represented in such ad images. Here, we examine how one major advertising platform (Facebook) delivers ads that include pictures of people of varying ages, genders, and races. We develop techniques to isolate the effect of these demographic variables, using a combination of both stock photos and realistic synthetically-generated images of people. We find dramatic skews in who ultimately sees ads solely based on the demographics of the person in the ad. Ads are often delivered disproportionately to users similar to those pictured: images of Black people are shown more to Black users, and the age of the person pictured correlates positively with the age of the users to whom it is shown. But, this is not universal, and more complex effects emerge: older women see more images of children, while images of younger women are shown disproportionately to men aged 55 and older. These findings bring up novel technical, legal, and policy questions and underscore the need to better understand how platforms deliver ads today.

ACM Reference Format:
Levi Kaplan, Nicole Gerzon, Alan Mislove, and Piotr Sapiezynski. 2022. Measurement and Analysis of Implied Identity in Ad Delivery Optimization. In *Proceedings of the 22nd ACM Internet Measurement Conference (IMC '22), October 25–27, 2022, Nice, France.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3517745.3561450

IMC '22, October 25–27, 2022, Nice, France

ACM ISBN 978-1-4503-9259-4/22/10…$15.00
https://doi.org/10.1145/3517745.3561450

1 INTRODUCTION

Advertising is now the primary way in which many internet companies fund their services, ranging from web platforms such as Google to social media sites such as Facebook and Twitter. To make their platforms attractive to advertisers, these companies often collect large amounts of data to infer various characteristics and interests of their users [33, 52, 66]. These inferences are then used to allow advertisers to precisely specify (i.e., *target*) their desired audience, and to allow the platforms to show (i.e., *deliver*) each ad to the subset of targeted users predicted to be most receptive. As a result, there has been extensive study and debate around the misuse of adverting platform targeting options [4, 38, 45, 56, 67], as well as more recent work demonstrating how advertising platform delivery algorithms can steer ads towards skewed subsets of the targeted users, based solely on the content of the ad [13, 14, 42].

But concerns around advertising are not unique to online ads. Since well before the creation of the Internet, advertising has been a key way in which people are informed about goods, services, viewpoints, and opportunities. Due to the ubiquity of advertising and the impact that it can have on individuals and society at large, governments have developed rules that regulate advertising in specific domains. For example, concerns over ways in which advertising could—either inadvertently or intentionally—reinforce historical inequities led to a number of U.S. civil rights laws [1, 2, 5] that regulate advertising for certain opportunities, including housing, employment, and credit. At the same time, purposefully created advertising can be used to counter historical inequities: companies aiming to increase the diversity of their workforce may choose to use images of non-male-presenting individuals or people of color in their recruitment materials. Such an approach has been shown to attract more applicants from historically underrepresented groups [18].

Due to both misuse [39] and legal settlements [3], advertising platforms are now removing particularly problematic targeting criteria, as well as limiting advertisers' ability to target small groups of users ("micro-target"). For example, on Facebook, ads for jobs, credit, and employment can no longer target users by gender, age, or race [40]. But these actions by the platforms have a side-effect: by removing targeting options, the power to choose which users ultimately see the ads is further shifted towards the platforms themselves, as decided by their ad delivery algorithms. Unfortunately, investigating these black-box systems remains a challenge.

In this paper, we examine how demographic information conveyed in ad images through the presence of diverse individuals influences the decisions made by ad delivery algorithms. We are concerned that the set of users who are ultimately shown the ad (called the *actual audience*) may be a skewed subset of the users who the advertiser targeted (called the *target audience*). As a concrete example, imagine you wish to advertise a job online, and you can

choose between a picture of a white or a Black person to include in the ad image (e.g., see the ads in Figure 1). We aim to understand how this choice influences who the ad delivery algorithm will show your ad to when everything else about the ad is held equal, including the users whom you target.

To do so, we focus on one of the largest online platforms (Facebook); we develop novel methodologies to measure skews of the actual audience by age, gender, and race, despite our lack of access to platform internals and the presence of numerous sources of noise. In brief, we became a Facebook advertiser, and ran over $2,800 of Facebook ads to test how the Facebook's ad delivery algorithm responds when ads are run with pictures of people from different demographic groups. While our results are limited to a single online platform (we were unable to study multiple platforms, as we do not have access to the necessary targeting and reporting tools on other platforms), we believe our results highlight the need for careful study of advertising platforms and delivery algorithms writ large.

Overall, our paper has the following contributions:

We demonstrate for the first time that ad delivery algorithms can deliver ads in a substantially different manner based *only* on the demographics of the person pictured. For example, ads containing stock images of women were delivered to an actual audience of 50% women, but this varies significantly by age: pictures of older women and female children are delivered primarily to women (58% and 55% women, respectively), whereas pictures of teenage women are delivered primarily to men (43% women).

We expand on previously-developed techniques for measuring the racial makeup of the actual audience by using a combination of voter records from multiple U.S. states [31, 51] and Facebook's Custom Audience [22] feature.

Using real-world pictures of people risks introducing other variables into the image, which the ad delivery algorithm may use when deciding how to deliver ads (e.g., the choice of clothing color, facial expression, etc). We address this challenge by developing techniques to synthesize carefully-controlled images using a deep-learning network [43], enabling us to create images of the same "person" that hold constant such variables and vary only by the demographics we study.

We show that Facebook's ad delivery algorithm responds almost identically when using ads that feature synthetic faces, demonstrating that the algorithm is indeed making its delivery decisions based on the demographic features and not other variables (e.g., choice of clothing). For example, ads containing synthetic images of adult Black people were delivered to 81% Black users on average, while ads containing synthetic images of adult white people were delivered to only 50% Black users on average.

Using a regression analysis, we isolate the independent effects of different demographic features in Facebook's ad delivery algorithm's decisions. We show that multiple features have a statistically significant role, including age, gender, and race.

Finally, we demonstrate the impact that these delivery decisions can have on "real-world" ads by running employment ads using our synthetically-generated faces. We show that the skews in delivery observed in our test ads persist when running "real-world" ads, albeit to a lesser (but statistically significant) degree.

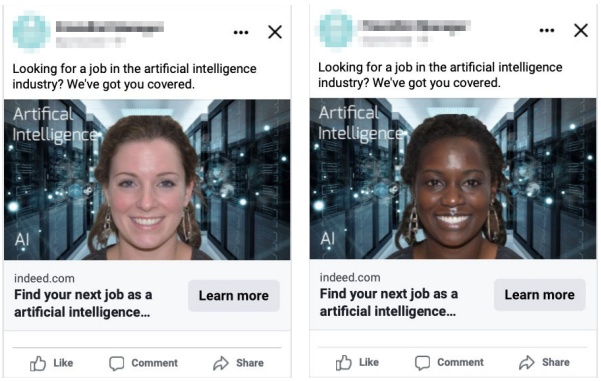

Figure 1: Two examples of job ads we ran. Despite being run at the same time, with the same budget, and targeting the same balanced audience, the ad on the left was delivered to 56% white users, whereas the ad on the right was delivered to only 29% white users.

Taken together, our results both underscore the power that ad delivery algorithms on social media platforms such as Facebook have today, and show how systems designed with neutral-sounding objectives ("delivering relevant ads to users") can inadvertently bake in unwanted bias. For ads in certain categories—notably housing, credit, and employment, which Facebook already has a separate advertising flow for—our findings raise questions about how existing civil rights protections in the U.S. may be implicated. Moreover, our findings call into question the approach of removing explicit features from machine learning training [58] or limiting targeting options from advertisers [3]; doing so increases the flexibility that ad delivery algorithms have when choosing the actual audience, and limits advertisers' ability to correct for any observed skews. Overall, our results further highlight the need for increased transparency in the advertising ecosystem, especially given the influences that the platforms they power have over end users' access to information.

The remainder of this paper is organized as follows: § 2 provides background on advertising platforms and advertising strategies, and details prior work. § 3 explains our methodology for creating and running ads, and measuring how they were delivered. We also elaborate on how to interpret the results of our statistical analysis. § 4 gives an in-depth discussion of the ethical concerns that our work brings up, and how we addressed them. § 5 describes our experiments with stock and synthetic images and the accompanying analysis, and § 6 describes our experiments with "real-world" ads in protected categories. § 7 lists the limitations of our approach and § 8 provides a concluding discussion. Appendix A discusses an additional experiment that involved controlling for poverty, finding similar trends to our main experiments. Finally, all ads along with their delivery statistics can be found on the project website: https://facebook-targeting.ccs.neu.edu/.

2 BACKGROUND

In this section, we provide background on how today's large-scale advertising platforms work, an overview of studies on how images of people are used in advertising, and a survey of prior work related to this study.

2.1 Advertising platforms

Major advertising platforms today, including Facebook, Twitter, and Google, are powered by *ad auctions*; these select which advertiser gets to show their ad based on the outcome of a virtual auction between advertisers. Here, we provide a brief overview to help explain the surprising complexity in the implementation of this process. At a high level, the functionality of major ad platforms can be broken down two phases: *ad creation* and *ad delivery*.

Ad creation The first phase is ad creation, where the advertiser submits their ad to the platform and makes choices about how they wish to have it delivered. In particular, the advertiser must:

(1) Upload the *ad creative*, which consists of the text, image, videos, and destination link that comprise the ad itself.
(2) Select the *target audience*, which is the subset of platform users eligible to receive the ad. We provide more details below.
(3) Choose the *objective* and *budget*, which specify what the advertiser is trying to achieve and how much they are willing to spend. Common objectives [7] on Facebook include Traffic (meaning the advertiser wishes to drive users to click on the ad and be directed to their website), Conversions (meaning the advertiser wishes to have users purchase their product or service), and Awareness (meaning the advertiser just wishes to show the ad to as many users as possible). The budget that the advertiser chooses usually covers the entire ad run; advertisers are not typically bidding on users individually. Instead, the advertising platform places bids on the advertiser's behalf in the ad auction based on a number of factors; this process is called *bid pacing* and is typically opaque to the advertiser [9].

Ad targeting To aid in identifying the target audience, the platform provides advertisers with a wealth of ways in which to target users; here we highlight two mechanisms. *First*, the advertiser can create boolean expressions over user attributes including demographics, interests, and behaviors [28, 37, 46] to specify who is in the target audience. In this approach, the advertiser does not know the identity of users in the audience, but instead relies on the platforms' inferences about the users. *Second*, the advertiser can provide the platform with the list of *personally identifiable information* (PII), such as names, phone numbers, physical addresses, or email addresses [10, 22, 54], thereby specifying precisely who is in the target audience. In this work, we rely on Facebook's implementation of this approach, called *Custom Audiences*. Note that it is possible to combine the two mechanisms, refining a Custom Audience by selecting only those users who have particular attributes.

Ad delivery The second phase is ad delivery, where the platform makes decisions about which users see which ads. As mentioned above, this decision is ultimately done via an auction, meaning whenever an "ad slot" is available (i.e., a user is browsing the site), the platform holds an auction among all of the ads where the user in question is in the target audience. However, the amount that each ad "bids" in this auction is dependent on a number of factors, among which the advertiser's budget is but one. For example, on Facebook, the amount that the ad "bids" in the auction is referred to as the *total value*, and it is calculated as [27, 36]:

$$\text{Advertiser Bid} \quad \text{Estimated Action Rate} + \text{Ad Quality}$$

In this equation, the Estimated Action Rate is Facebook's estimated probability that this particular user will help the advertiser achieve their objective, and the Ad Quality is a measure of whether the ad is scammy, clickbait, or contains low-quality images [27]. Importantly, the calculation of Estimated Action Rate is done via machine learning [36], raising concerns about whether pre-existing societal inequities and biases could inadvertently be reflected in its estimates, thereby skewing delivery. As described below, recent work has demonstrated that this is indeed the case for ads in certain contexts [13, 14, 42].

Reporting Finally, ad platforms provide feedback so that advertisers can monitor how their ads are performing. Typically, these reporting features will include information about the actual audience, including the number of *impressions* (how many times the ad was shown), the *reach* (how many unique users the ad was shown to), and demographic breakdowns of whom the ad was shown to (e.g., the number of men, women, and users of unknown gender, age distribution, or the locations where the ad was delivered). Importantly, ad platforms typically do not tell the advertisers precisely *which* users were shown the ad, or which users clicked on the ad (though advertisers can often use other approaches—such as using first-party cookies or syncing with data brokers—to identify users who visit their own website). In Section 3, we extend prior techniques [13, 14] to allow us to infer additional demographics of the actual audience beyond the ones that Facebook reports.

2.2 Related work

We now provide a brief overview of work related to this paper, covering work that focuses on representation of people in advertising and studies of real-world advertising systems.

Representation in advertising Researchers have long tracked the representation of different demographic groups in advertising and studied the impact this representation has on both the majority and minority groups. Historically, minority groups were underrepresented in advertising [15, 64]. Furthermore, even their scant portrayal has often been limited to certain roles [20] and perpetuated race and gender stereotypes [21]. Such enforcement of stereotypes in job ads has been shown to discourage counter-stereotypical candidates from applying [19, 32]. Over the years, the minority presence in advertising has increased and has been shown to bring about a range of positive effects. Job advertisements that feature more diverse individuals elicit higher interest from minority candidates [41, 53] without discouraging majority candidates from applying [18]. Similarly, brands which increase the minority presence in their product advertising see increased awareness and engagement among minority customers [16]. Often, exposure to diverse individuals in advertising and other media can have even more tangible effects on those who see themselves represented. For example, Good et al. showed that including counter-stereotypical imagery in science textbooks measurably increases the performance of female students without negatively affecting the performance of male students [35]. Nowadays, it is common for companies to signal that they support diversity through advertising, even if the organizations are not well fit to retain minority employees [47].

Legal and policy rules As a result of these studies and others, there are growing concerns around discrimination in targeted advertising regardless of whether the skew was caused by direct human input or an algorithm [55]. In the U.S., existing regulation such as the Fair Housing Act [2], Section 704b of Title VII of the Civil Rights Act [5], and Section 230 of the Communications Decency Act [6] are often interpreted in tandem to tackle this issue [23]. In particular, indicating preference for race, gender, sexual orientation, religion, age, and other protected classes in housing and employment ads could violate these laws. While the content of ads, such as the person being pictured, may not be a clear indication of advertiser bias, when platforms are skewing the delivery of such ads—thereby withholding them from a not-pictured group—it could reveal a preference that has been deemed illegal in other contexts [24].

Bias in online advertising There is a body of work that has demonstrated how the targeting tools provided by advertising platforms can be abused for malicious purposes, such as voting manipulation [25], exclusion from housing opportunities based on inferred ethnicity [17], or from employment opportunities based on age [44]. Platforms have attempted to address such abuse by limiting these targeting options in sensitive contexts, for example by enforcing minimum target audience sizes [63], or removing targeting options that explicitly mention a protected class [26, 30]. In fact, Facebook was sued in 2018 by the National Fair Housing Alliance, who alleged that Facebook's targeting options enabled violations of the Fair Housing Act [3]; as part of the settlement of this lawsuit, Facebook created a separate ad creation flow for housing, credit, and employment ads [40] and removed a number of targeting options from such ads [8]. These limitations mean that the ad delivery algorithms that determine which ads are actually shown have *more* power to decide who actually sees ads.

When designing ad delivery algorithms, platforms make the choice in a way that optimizes for both the advertiser's stated goal (for example maximizing clicks or conversions), as well as their own bottom line. For example, ad delivery algorithms have been shown to lead to gender and race skews in the delivery of ads [13, 42] as well as price discrimination, and echo chamber effects in the delivery of political ads [14]. The popular press also reported on unwanted effects that originate from ad delivery optimization. For example, when Musical.ly—now known as TikTok—was entering the U.S. market, they ran ads featuring young women. Facebook delivered these ads disproportionately to middle-aged men, presumably because that group had the highest engagement with such content [62]. This optimization for engagement has also been leveraged by scammers, who rely on the delivery algorithm to identify the users who are most likely to fall for them [61]. Such problems can be difficult to detect and measure because the harms may affect only relatively small fractions of disadvantaged users [12]. Nevertheless, prior work has not explored how images of people are treated by the delivery algorithm, an important topic given the frequent use of such images.

3 METHODOLOGY

In this paper, we aim to investigate how gender, race/ethnicity, and age—implied through the use of images of diverse people—influence the delivery of ads that feature such images. To do so, we use and extend a number of previously-proposed techniques for measuring ad delivery, and make novel use of existing tools for generating synthetic images of people.

3.1 Selecting images

We describe below the two sources of images during our experiments: stock photographs and synthetically-generated images.

Stock images We purchased stock images from a popular stock photo website, Shutterstock [60], and paid for a license to use them in advertisements. We chose the images to be a balanced set of images across estimated ages (child, teenager, adult, middle-aged, elderly), genders (male, female), and races (white, Black), using Shutterstock's search functionality to find candidate images. We annotated the images with the corresponding demographic labels manually. We then compared our labels with the text descriptions provided by the owners of the pictures whenever available. For each of the 20 possible combinations of these attributes, we selected five Shutterstock photos of different people; thus, we purchased the right to use 100 separate images. We provide a copy of the images, as well as information on their cost and licensing on the project webpage: https://facebook-targeting.ccs.neu.edu/.

Synthetic images While the stock images provide a realistic proxy of the photos an advertiser might use to promote a real-world product or service, they may also introduce noise into our measurements, as they vary in composition, head positions, lighting, facial expressions, backgrounds, clothing, etc. To isolate the effect of the demographic characteristics of interest from the spurious effects of such noise, we need images where we can control all variables. To do so, we use the StyleGAN [43] deep-learning-based framework for generative image modeling. Given any 512-element input vector, StyleGAN will produce a 1024 1024 px "headshot" of a person; see Figure 6 for examples. Importantly, these are *not* images of real people but merely an output of a deep learning system trained on images of real people. In Section 5.4, we describe the technique we used to identify the latent directions for the demographic characteristics we consider in this paper (age, gender, race). Modifying the activation values of the synthesizer network along these latent directions allows us to change one demographic attribute of a particular synthetic person at a time while holding other aspects of the image constant.

3.2 Running ads

We now describe how we run ads on Facebook and measure the demographics of the resulting actual audience.

Ad setup In order to measure whether Facebook's ad delivery algorithms are introducing skews in the actual audience along the demographics we study, it is important that we isolate the impact of the algorithm from other effects (e.g., the decisions of other advertisers, the relative user activity levels, etc). To this end, we re-use previously published approaches to running ads [13, 14, 42] while controlling for these effects. Unless otherwise specified, when running a set of ads, we always launch all ads at the same time,

from the same advertising account, targeting the same audience,[1] with the same budget, and with the same ad creative features other than the image (e.g., the ad headline, text, destination link, etc). For all experiments, we created ads through the Facebook Marketing API, set the ads to all have the same daily budget (between $2 and $3.50, depending on the experiment), used the objective of Traffic (consistent with prior work [13, 14]), and ran the ads for exactly 24 hours. Thus, for a given experiment, the *only* difference between the ads is which image we choose to include. Table 2 shows details about the campaigns discussed in this paper, all campaigns were run from an ad account created in 2019, except for the "real-world" ads, which were run from an account created in 2007.

Balancing audiences Selecting the target audience is a part of the ad creation; we aim to create audiences that are as closely balanced as possible by the demographics we consider. To do so, we build Custom Audiences based on publicly available voter records, similar to prior work [13].[2] Recall that Custom Audiences allow the advertisers to target lists of particular individuals with identifiable information known to the advertisers. For example, an advertiser can create a Custom Audience from a list of names and postal addresses, and then run advertisements that are shown to only those Facebook users [22]. We select names and addresses from the voter records, using them to create Custom Audiences.

We sample voter records in a stratified way such that age, gender, and race are not correlated. For each age group: 18–24, 25–34, 35–44, 45–54, 55–64, 65+,[3] we select voters such that the number of men and women is equal, as is the number of Black and white voters, and as are the intersections of race and gender. For example, there are as many white men as Black women from each state. We repeat the process in separately for each age group. Table 1 details the size of each of these lists. Doing so ensures that an equal number of users of each demographic are selected in our target audience, but does not imply that the actual audience needs to be evenly divided in the same manner: each demographic group may not have the same percentage of voters with Facebook accounts, may not have the same level of Facebook activity, and may not be equally "priced" based on the targeting of other advertisers. Thus, when analyzing our results, we only compare ads that experienced the same running environment, looking for how two ads that differed *only* by the demographics of the person in the image were delivered by Facebook.

3.3 Measuring delivery

After we have run a given ad, we wish to measure the demographic breakdown of the actual audience along the lines of age, gender, and race. Measuring the first two of those is much easier than the third.

Measuring age and gender Recall that as a Facebook advertiser, we have access to Facebook's marketing tools. Whenever we run an ad, Facebook provides detailed statistics on how our ad is being delivered, showing the number of ad impressions for users with different attributes. We are able to access this data through Facebook's Insights API, which gives us a breakdown of the actual audience based on gender and age, providing us with a direct mechanism for measuring these attributes.

Age range	Group size	Total
18-24	44,968	179,872
25-34	53,586	214,344
35-44	51,469	205,876
45-54	61,893	247,572
55-64	68,211	272,844
65+	78,719	314,876

Table 1: Breakdown of the number of voters selected from each state within each combination of race (white, Black) and gender (male, female). The Total column represents the total size of the target audience in each age range.

Measuring race To measure the racial makeup of the actual audience, we build upon a methodology developed in prior work [13]. We first select two locations that are physically far apart (call these locations A and B). We then create a Custom Audience from the names and addresses in the voter records, selecting only white voters from A and Black voters from B (specifically, an equal number of each). We also create a "reversed" copy of this Custom Audience, with Black voters from A and white voters from B. We then run *two* copies of the ad, one targeting each of these two Custom Audiences independently. Once the ads start delivering, Facebook's marketing tools tell us the breakdown of *where* the ad is being delivered. For the ad targeting the first Custom Audience, we know that the only individuals we targeted in location A were white, and therefore we can count every delivery to location A as delivery to a white person (and conversely, every delivery in location B as delivery to a Black person). We can make a similar, reversed, inference for the second ad.

In this paper, we use the states of Florida and North Carolina as our two locations. These two states both have publicly-available voter records with self-reported race in them, and they are sufficiently far apart to minimize any error from users who travel from one state to the other. An overview of this entire methodology for measuring the race of the actual audience is shown in Figure 2.

Discussion There are a few aspects of this method that are not intuitive, yet crucial to understand. *First,* our method does not rely on existing racial imbalances in the locations we pick. Following our example, location A does not have to be majority white and location B does not have to be majority Black. We specifically select individuals within these locations whose self-reported race we know, balancing the target audiences so they are equal regardless of the overall community makeup. *Second,* in all experiments, we run two copies of the ad in parallel to "reversed" Custom Audiences. In our analysis, we aggregate both copies (i.e., to calculate what fraction of delivery to white users, we add the number of deliveries to location A in the first copy and the number of deliveries to location B in the second copy, and divide by the total). This way, we minimize the influence of any confounding non-race related differences between the two locations we chose. *Third,* one concern is that users who are

[1] Other than the region-based proxy split, where we run two copies of every ad, targeting each of the two region splits.
[2] We recognize the voter information might not always be fully accurate and current. Nevertheless, we expect that such errors to be infrequent, and to not be heavily biased towards a single demographic.
[3] We use these age ranges as they are the same ones used by Facebook in their marketing tools for reporting breakdowns of delivery.

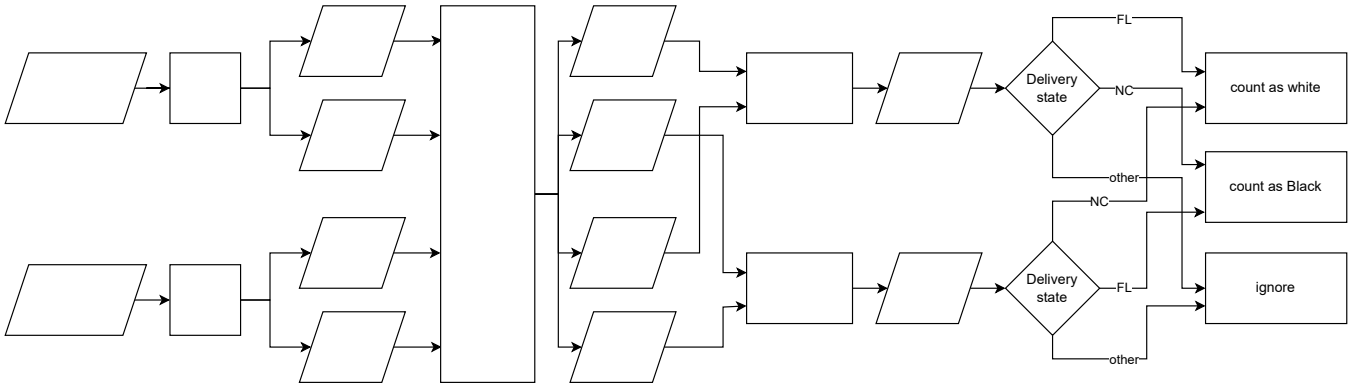

Figure 2: Flowchart detailing how we use Facebook's Custom Audiences feature to measure how our ads are delivered along racial lines.

traveling between our two locations may be miscounted. It is not possible for us to measure how much noise *exactly* this introduces to our measurements, but we take steps to minimize it. We chose two non-adjacent states as the locations to select voter records from (Florida and North Carolina, requiring at least five hours car travel). We expect that the number of individuals traveling between the states would be small compared to the numbers of those who stay within their state of residence. Moreover, since we run two parallel copies with the race-state assignment reversed, error in the measurement should be treated as noise rather than bias: any traveling by a person of either race will affect the results in the same way.

We note that Ali et al. [13] reported over 10% of their impressions fell outside of their target Designated Market Areas (DMAs) [29]; with our approach that uses states rather than DMAs, that fraction drops to less than 1%. Since this means that the fraction of users who are traveling to *all other 48 states combined* represents less than 1% of the total delivery, the fraction of users who are traveling to the specific other state is likely to be much smaller. This observation is in line with human mobility research that shows that much of day-to-day travel is contained within smaller, meaningful areas [11].

3.4 Interpreting linear regression results

Throughout the paper we rely on linear regression to measure the effects of implied demographics on the demographic makeup of the actual audience. Table 4 shows the coefficients that result from the linear regression analysis for different problems. One can interpret the coefficients in the following fashion. The intercept is the mean value of the target (dependent) variable when all the explanatory (independent) variables are equal to 0. For example, in Table 4a the intercept for the "% Black" model is 0 5697. This means that 56.97% of the actual audience is Black when all the explanatory variables are 0, i.e. when the person presented in the image is a white adult male. The coefficient of each other variable describes how much the target variable changes when that variable increases by 1 (or, in our case, becomes true), while other variables are held constant. For example, in Table 4a the coefficient of the variable "Black" for the "% Black" model is 0 1812. This means that a picture of a Black person will reach an actual audience where the fraction of Black users will be 18.12 percentage points higher than if a person shown in the image was the same gender and age, but white. These coefficients are additive: to estimate the fraction of the actual audience that is 65+ years old for an image of a white elderly woman, we would add the corresponding coefficients: intercept, female, elderly. Note also that coefficients are marked with their statistical significance: 0 05; 0 01; 0 001; a lack of symbols indicates that the coefficient is not statistically significant.

Finally, we provide the ^{2} score for each model we train. The number is the fraction of variance in the data that can be explained by the model. An ^{2} value of 1 means that the model can perfectly predict the target variable using the explanatory variables, while a value of 0 means that the features offer no explanatory value.

4 ETHICS

Our methodology and experiments bring up a number of ethical issues, and we provide more details below about how we addressed them. Our study design has been reviewed by our organization's Institutional Review Board and deemed exempt (Northeastern University IRB Decision #18-11-13). Importantly, the user data we use in the paper—voter records from Florida and North Carolina—are public records by both states' laws [50, 65].

4.1 Harm minimization

We carefully considered the impact and potential harm that our experiments may have on people whose likeness we used in images, Facebook users, Facebook content moderators, and Facebook itself.

For *people whose images we used in ads*, we obtained the images from a stock photography website, Shutterstock. We paid for the right to use these images in advertising (as would any other advertiser), and we only chose images of typical "headshots" (the images themselves are provided on our project webpage).

For *Facebook users*, the potential harms come primarily from being shown our ads. We minimized any such harms by only running ads for actual opportunities, and we made sure that the ad content matched the linked site (i.e., we only ran "legitimate" ads, in the sense that the destination link was a website that was relevant to the content of the ad). We did not collect any information about the users who clicked on our ads, as they

#	Ads	Age-limit	Images	Date	Length	Reach	Impressions	Spend	Section
1	200	No	Stock	Apr 05, 2022	24 hours	24,248	36,535	$ 387.59	§5.2
2	200	Yes	Stock	Mar 30, 2022	24 hours	34,480	80,758	$ 686.87	§5.3
3	200	No	Synthetic	Mar 3, 2022	24 hours	27,192	44,911	$ 386.67	§5.5
4	88	No	Synthetic, with job background	May 12, 2022	24 hours	18,356	22,090	$ 216.71	§6

Table 2: Overview of the ad campaigns that we present in this section. Shown are the source of each ad campaign's images, how long it ran for, the number of times it was shown (Impressions), and the number of unique users it was shown to (Reach).

did not visit websites under our control (e.g., our job ads linked to pages on indeed.com, a job-hunting site). While we do use images of children in some of our ads, we do not advertise to children. In fact, by the virtue of using Custom Audiences of only eligible voters (aged 18 and older), we also prevent any accidental displaying of our ads to children due to algorithmic optimization.

For *Facebook's content moderators*, the harms come primarily from reviewing our ads as part of their content moderation role. We minimized such harms by not creating or running any ads which could be seen as distressing.

For *Facebook itself*, the potential harms come primarily from our activities as an advertiser on their platform. We minimized any infrastructure and financial burden to Facebook by only using their official advertising APIs and collecting the delivery data from a single vantage point without parallelizing queries. We paid for the ads that we ran on the platform, as would any other advertiser. Finally, for the ads we ran for protected categories (e.g., employment), we always flagged our ads as being in these categories as part of Facebook's Special Ad Categories flow [40].

4.2 Notions of gender, and race

Throughout this paper we primarily examine three demographic axes: age, gender, and race. The latter two of these require careful consideration. For race, we use voter records from Florida and North Carolina as our ground truth; in both of these states, voters self-report their race when they register to vote. Both of these states limit [31, 51] the available race options to those used by the U.S. Census.[4] As such, our analysis inherits biases present in this data collection, both from voters who may not wish to self-report their race, and from voters for whom the available race options do not accurately capture their view of their race. Both North Carolina and Florida allow only self-reported gender options of Male, Female, and Unknown. Similarly, Facebook only reports gender as Male, Female, and Other. Thus, we inherit any biases affecting voters and Facebook users whose gender identity is not accurately represented among these options.

Furthermore, we refer to demographic information hinted at in the synthetic pictures as "implied" demographics. We make this distinction to avoid conflating self-reported demographic information of real individuals with stereotype-driven pixel perturbations. We construct these images such that a machine learning library classifies their gender or race according to our hints, but do not make any statements pertaining to the accuracy of Deepface's demographic inferences.

[4] Available options are: American Indian or Alaskan Native; Asian Or Pacific Islander; Black, Not Hispanic; Hispanic; White, Not Hispanic; Other; Multi-racial; and Unknown.

5 RESULTS

We now detail the experiments we performed and their results. We begin by providing an overview of our results (§ 5.1), and then describing our experiments with stock photos (§ 5.2 and § 5.3), followed by our approach to generating synthetic images (§ 5.4), and our results with the synthetic images (§ 5.5). Finally, we describe results from "real-world" ads in protected categories (§ 6).

5.1 Experiments overview

Throughout this section, we run a number of experiments to explore how Facebook's ad delivery algorithm reacts to ads with images of people from different demographics. In each experiment, we run a number of ads (called a *campaign*) with images of different people; unless otherwise noted, all other aspects of the ads in a given run are the same. To help provide a quick reference throughout this section, Table 2 provides an overview of all of the campaigns we ran and in which section the results are discussed.

5.2 Stock images

Recall from Section 3.1 that we selected 100 images from Shutterstock, balanced across age ranges (five age "buckets"), gender (male, female), and race (white, Black).

Ad campaign overview We first investigate the delivery of stock photos, using the methodology described in Section 3.2 to gather delivery data for race in addition to the data on age and gender delivery Facebook automatically provides. In each ad we use a different stock photo as the ad image, unedited except for cropping the image into a square around the person's face. We adapt the ad text from a real ad that our university had previously used for one of its professional Masters programs. We replace any mentions of the university in the ad text but keep the phrasing, call to action, and destination link (to a real website for a project management career guide published by the university). This ensures that any interested viewers of our ad will be able to access the advertised content the same as with any other ad, and will not be misled by the content.

We ran 200 versions of this ad at the same time, all from the same account and with the same budget; this is referred to as Campaign 1 in Table 2. Specifically, for each of the 100 images, we ran one copy targeting an audience of white users from Florida and Black users from North Carolina, and another copy targeting the reverse audience. A total of 306 impressions (0.8%) occurred outside of

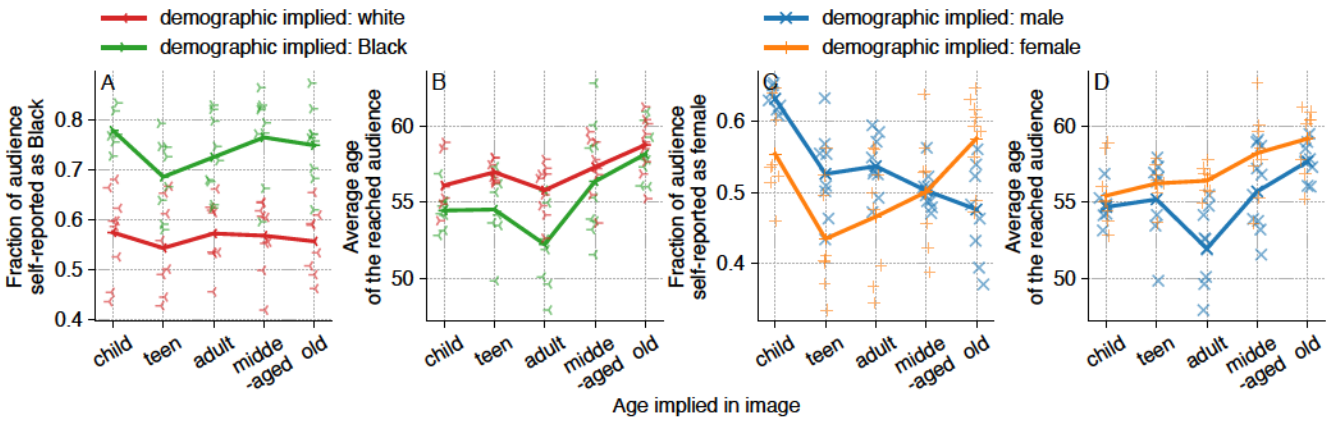

Figure 3: Delivery statistics of ads featuring stock images. A) ads with Black people are delivered more to Black users compared to ads showing white people. B, D) Images of older people of both races and genders tend to be delivered more to older users. C) Images of children are shown more to women; images of teenage and adult women are shown more to men.

Florida and North Carolina (presumably these occurred due to Facebook users who were traveling or had moved out of these states); we disregarded these impressions in the resulting analysis.

Aggregate results We begin our analysis by providing an overview for how these ads were delivered. Table 3 provides a high-level overview of how these 200 ads were delivered, aggregating ads with images with different implied identities in different rows, and showing the breakdown of the users to whom the ad was shown along the columns. We make a number of observations.

First, focusing on the first column of results, we observe that images of Black people are delivered more to Black Facebook users (73.8%) when compared to images of white people (56.3%). Similarly, images of teenagers are also delivered less frequently to Black users (61.4%) when compared to other images of other age groups (between 65.1% and 66.4%). Recall that the only difference between our ads is the choice of image, meaning these differences in delivery are due to the delivery algorithm rather than any targeting choices.[5] *Second*, focusing on the second column of results, we observe that images of children are delivered more to women (59.4%) than images of other age groups (between 48.2% and 52.4%). *Third*, while images of children, teenagers, and adults are delivered to between 70.5% and 75.6% older Facebook users, this percentage increases for images of middle-aged or elderly people (78.2% and 80.5%, respectively).

Implied identity	Demographics of actual audience		
	% Black	% Female	% Age 45+
Race			
Black	73.8%	53.0%	78.9%
White	56.3%	50.8%	72.2%
Gender			
Male	65.4%	53.2%	72.4%
Female	64.1%	50.5%	78.6%
Age			
Child	65.1%	59.4%	72.5%
Teen	61.4%	48.2%	75.6%
Adult	65.1%	50.5%	70.5%
Middle-age	66.4%	50.2%	78.2%
Elderly	65.8%	52.4%	80.5%

Table 3: Delivery breakdowns of stock image experiments when targeting all ages and optimized for Traffic. Implied identity of the person in the ad image affects who the ad is shown to.

Detailed results The aggregate results presented in Table 3 could obscure trends that exist for intersectional groups (e.g., images of *female* children being treated differently that those of *male* children). To explore these in more depth, Figure 3 presents a detailed breakdown of how each individual ad image was delivered. For each graph, the x-axis varies the age of the person in the image, with the two colors represent images of different genders or races; the y-axis shows the demographics of the actual audience. Each individual ad image is represented by a tick mark, and the average across all images with the same demographic is shown by a line.

Focusing on graph Figure 3A, we can immediately see that the aggregate difference between images of Black people (green arrows pointing right) and white people (red arrows pointing left) in delivery to Black Facebook users persists across images of people of all ages. In fact, the delivery can almost be cleanly separated, with almost all images of white people being delivered more to white Facebook users than almost all images of Black people.

However, the trends are more complex for other demographics. Consider the graph in Figure 3C, which shows how images of men (blue exes) and women (orange crosses) are delivered to Facebook users of different genders. We first observe that for both genders, images of children are delivered more to female Facebook users. After this, the two lines change behavior. For images of women, images of teenage women are delivered much more to men (56.6%) than any other female age group; in fact, as the age of pictured

[5]Additionally, note that we do not necessarily expect a 50%/50% delivery to white/Black users, as Facebook users in different demographics may exhibit different levels of online activity and may be targeted differently by other advertisers (making their relative costs different). Instead, we focus on comparisons between ads run within the same experiment, as all such ads experienced any such effects equally.

women increases, they are delivered increasingly more to women. This is in line with the press reports about images of teenage women delivered predominantly to men [62]. For images of men, there is a trend where as the age of the pictured man increases, the more likely it is to be delivered to men.

Finally, in Figure 3B and Figure 3D, we show the average age of the Facebook users to whom the ad was delivered, based on the race (Figure 3B) and gender (Figure 3D) of the person in the image. We observe an overall increasing trend that images of older people are shown to older Facebook users, with one notable exception: images of adult men, which are disproportionately delivered to younger users (e.g., only 64.6% of images of adult men are delivered to users 45 and older, compared to 76.5% of images of adult women).

Regression The results discussed above reveal multiple, intersecting trends, and bring up questions about the role that each demographic feature plays independently. To separate the effects of different demographics in the images—and obtain a measure of statistical significance—we employ *linear regression* analysis. In brief, a linear regression is a way of modeling how multiple *explanatory features* (the demographics of the people in our ad images) can be combined in a linear fashion to explain the variance in a *target variable* (the gender, race, or age makeup of the actual audience). The output of a linear regression provides estimated coefficients for each explanatory feature along with a measure of the statistical significant of that coefficient.

To set up our regression, we represented the demographic information (gender, race, and age) in each ad image as dummy variables[6] and used these binary values as features into the linear regression model. We create three different models with the same features (independent variables) but different target (dependent variable): fraction users in the actual audience that are Black (% Black), % Female, and % Age 65+.

The results of our regression are shown in Table 4a, and we make a number of observations. The R^2 (R-squared) value, shown in the last row, indicates how much of the variance in the target can be explained with the explanatory features. We see that the % Black delivery model explains 62.2% of the variance, while the corresponding models for % Female and % Age 65+ explain 26.2% and 46.4% of the variance, respectively. This indicates that that the demographic information contained within the image indeed does explain a large portion of the differences in the demographic makeup of the audiences of different ads.

Moreover, closely examining the coefficients of the resulting models (middle seven rows), we see that different features are statistically significant in different models. For delivery to Black Facebook users, we observe that the only significant feature is whether or not the image is of a Black person. For delivery to female Facebook users, we observe that the only significant effect is whether or not it is an image of a child. Finally, for delivery to elderly users, we observe that three variables are statistically significant: in decreasing order, whether it is an image of an elderly or middle-aged person, or whether it is an image of a woman.

[6]Given N possible values of categorical features, dummy encoding uses $N-1$ binary variables while one-hot encoding uses N binary variables. For the purposes of regression analysis dummy encoding is sufficient and the "missing" variable is assumed true when all others are false.

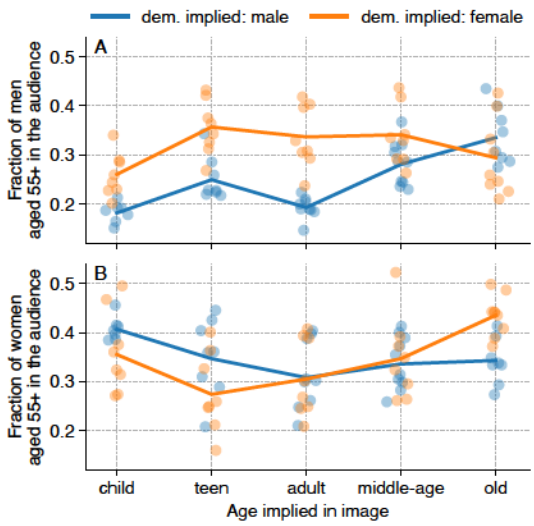

Figure 4: A) Older men receive many more ads depicting younger women than those showing younger men. B) The effect on older women is weaker and does not extend to images of adult/older men.

Contrary to our expectations, images of women are not delivered more to female users, as shown by a non-significant coefficient of the feature Female in the % Female model. As we show in Figure 4A, this effect can partially be attributed to the fact that men above the age of 55 receive a disproportionate number of ads depicting younger women.

5.3 Additional testing

One observation from the previous experiment is that, across the board, our ads were shown disproportionately to older Facebook users (e.g., over 70% of all ads were delivered to users ages 45 and over, as shown in Table 3; despite only 58% of our target audience being from this group, as shown in Table 1). Thus, we wanted to explore whether the same trends held if we prevented Facebook from delivering to such users. To do so, we utilize a feature on Facebook's advertising platform that allows us to limit the age of users in our target audience. We re-ran the previous experiment consisting of 200 ads, but limited the age of the our target audience to be 45 or younger; this is referred to as Campaign 2 in Table 2.[7]

We performed a similar regression analysis on the results of these ads as before, with the only change being the third target variable changed from % Age 65+ to % Age 35+ (since we artificially capped the maximum age of the actual audience to be 45, the oldest age group is now those aged 35-45 instead of those 65 and older). The results of this regression are presented in Table 4b. Overall, we observe very similar delivery characteristics to the experiment that included to all ages, suggesting that our results are not an artifact of the makeup of our target audience, nor due solely to the large fraction of delivery to older users. Notably, the results for this experiment are in many cases *stronger* than our previous result: the model where the target variable is % Black has a slightly higher R^2,

[7]Due to delays in getting these ads through the ad review process, they were launched with a somewhat higher per-ad budget ($3.50) compared to the previous ads ($2.00).

	a) Stock Images			b) Stock Images (younger users)			c) StyleGAN Images		
	% Black	% Female	% Age 65+	% Black	% Female	% Age 35+	% Black	% Female	% Age 35+
Intercept	0 5697	0 5030	0 3286	0 5520	0 4386	0 4433	0 5480	0 3714	0 4733
Black	0 1812	0 0258	0 0028	0 2534	0 0185	0 0343	0 2344	0 0212	0 0169
Female	0 0278	0 0258	0 0359	0 0146	0 0780	0 0362	0 0044	0 1377	0 0134
Child	0 0281	0 0924	0 0328	0 0829	0 1328	0 0888	0 0260	0 1643	0 0917
Teen	0 0315	0 0205	0 0224	0 0094	0 0301	0 0240	0 0098	0 0362	0 0644
Middle-aged	0 0217	0 0020	0 0508	0 0259	0 0155	0 0459	0 0136	0 0102	0 0076
Elderly	0 0077	0 0235	0 1180	0 0511	0 0274	0 0044	0 0480	0 0111	0 0402
^{2}	0 622	0 262	0 464	0 638	0 314	0 467	0 606	0 496	0 225

0 05; 0 01; 0 001

Table 4: Linear regression results for stock photos to Facebook users of all ages. Shown are three separate models (columns), each with different target variables. Images of Black people are statistically shown to increase delivery to Black people, while images of children increase delivery to women, and images of older people and women increase delivery to the elderly.

and a much larger coefficient for images of Black people. Further, we note that when we limit the maximum age of the targeted audience, women do receive more ads that feature women.

5.4 Generating synthetic images

Following these two experiments, our next goal was to separate out any possible effects from image composition, background, or lighting that could be contributing to the differences in delivery we observed. In other words, we wish to demonstrate that it is indeed the features of the *person pictured* that is leading to the delivery differences we see, rather than other features of the image that Facebook's ad delivery algorithm may be picking up on. To do so, we leverage the StyleGAN 2 [34] deep learning-based image generation tool.

However, we are faced with a challenge: we wish to generate images of a synthetic person where we can manipulate their age, gender, and race while holding all other features of the image constant. To do so, we need to determine how the different demographic characteristics are represented in the neural network. We follow and extend the steps laid out by Nikitko [49].

Generating and classifying face images We first generated 50,000 random face images by providing StyleGAN 2 with 50,000 input vectors each with 512 random values. When applied to StyleGAN 2, each input vector activated neurons of the network in the processes referred to as *mapping*. We saved the activation values for each neuron in each layer of the network and represented them reshaped as a one dimensional vector. The network has 18 layers of 512 neurons each, so the resulting one dimensional vector has 18 512 9,126 values. We also *synthesized* and saved the actual face image.

We then used the Deepface [59] library to obtain the machine estimation of the gender, race, and age of the "person" in the image.[8] The library supports binary gender labels (male, female) and a number of race/ethnicity labels (White, Latino Hispanic, Middle Eastern, Black, Asian, Indian). Gender or race are constructs that cannot be read from an image of a person's face, let alone an image of a person "that" does not exist. However, for this paper we want to imply demographics to another machine learning algorithm, rather than make claims about any person's gender or race.

Finding the latent directions We now have two items of interest for each face image: the 9,126-length activation vector and the demographic labels (male, female, white, etc). The question we want to answer is: Given a face image, how can we perturb the vector so that it has more or less of a given demographic feature? In other words, if I have an image of a young person, how can I modify the activations to make the person appear older without changing their other characteristics?

To do so, we calculate the latent directions in the activation space that correspond to each demographic of interest. We determine these directions by performing logistic regressions with node activation levels as independent variables and the predicted characteristics as dependent variables. The fitted coefficients of the regression model are precisely the vector in the activation space that represents the direction of change.

In more detail, we create a single logistic regression model for gender with female as target. Further, we create a separate logistic regression model for each race/ethnicity as target and white as a distractor. Finally, we create a linear regression model with age as the target.

Once the latent directions are established, they can be used to move through the latent space and create images which differ by the requested feature, while minimizing changes to the background, clothing, and face position. To see an example of the images that this technique produces, see Figure 6. It is important to note that this approach is subject to all biases that arise from the combination of biases in self-presentation, training data, latent space allocation, and classification biases of Deepface. For example, changing the "gender" of a picture from male to female also tends to introduce a more pronounced smile.

5.5 Synthetic image results

We now explore how ads containing these StyleGAN 2 images are delivered. We first select five source face images, and then use the technique we just described to generate 20 images of the same "person", varying their age, gender, and race in the same groupings that we used when selecting our stock images. As before, we end

[8]Note that the Deepface library was developed by Facebook but we do not know whether it is used internally for the purposes of distinguishing human features.

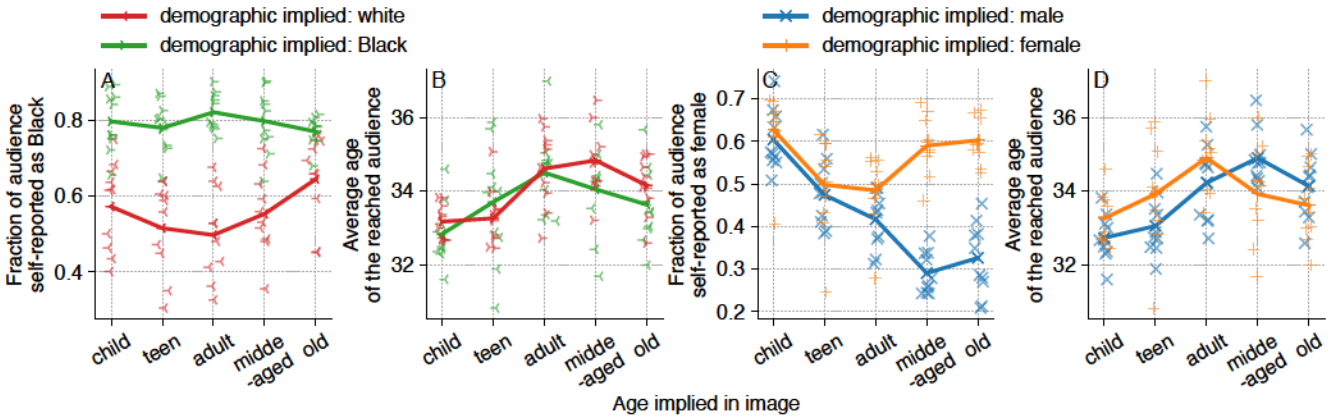

Figure 5: Delivery statistics of ads featuring StyleGAN images, revealing similar trends to those with stock images in Figure 3.

up with 100 total images, with five images in each combination of demographics.

We then run the same ads as we did in the previous experiments using these images, targeting the same age-limited audience (44 and under); this is referred to as Campaign 3 in Table 2. We collected and analyzed the results in the same manner as the previous experiments.

Figure 5 presents a detailed look at how these images were delivered. We can immediately observe that most of the trends that we observed on the stock photos persist when we use our synthetically generated faces: in panel A, images of Black faces are delivered significantly more to Black users; in panel B, images of older faces tend to be delivered to older Facebook users; and in panel C, images of female and male faces of different ages are delivered very differently, most notably with images of young women be delivered disproportionately to men.

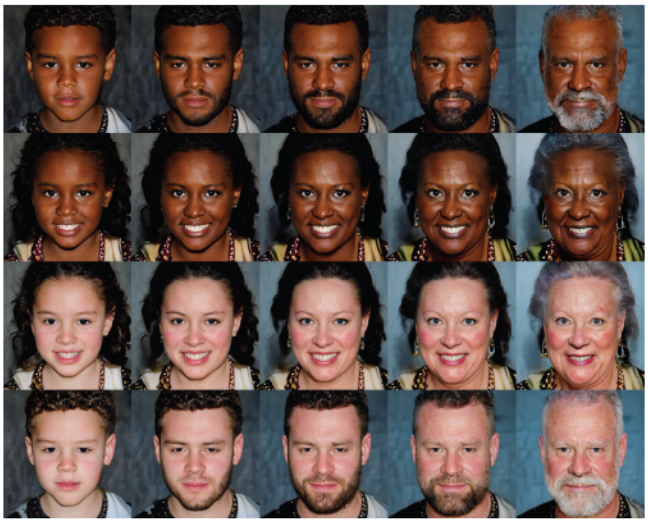

Figure 6: Images of faces generated using StyleGAN 2 by sweeping through the latent directions of gender, race, and age.

As a final point of analysis, Table 4c presents the results of a linear regression on these results. When compared to the previous regression for stock photos in Table 4b, we first observe that the R^2 and coefficients for the % Black model match the stock photo experiments extremely closely. This further cements our result that images of Black faces are indeed delivered more to Black users, everything else held equal. When focusing on the % Female model, we first note that the the R^2 is higher (0.496) than with the stock photos (0.314); while we do not know the source of this improvement in explainability for the stock photos, we hypothesize that it may be due to the biases present in Deepface that were carried over into our latent dimensions. Regardless, we observe that that the two independent variables that were statistically significant in the stock photo experiment continue to be significant here. Finally, focusing on the % Age 35+ model, we see the opposite trend, with the explainability going down relative to the previous experiment.

Note that we do not provide a statistical measure on the similarity of the models resulting from stock and synthetic images. The experiments were run at different times and thus were subject to different extraneous conditions. We present the two models side-by-side to show that the effects persists, but comparing the exact effect sizes is not appropriate.

6 REAL-WORLD ADS

As a final point of analysis, we explore the extent to which the skews induced by the ad delivery algorithm that we observed in the prior section impact "real" ads. Because "real" ads often have other features in the image beyond just an image of a face, we want to see whether these skews persist when such other features are present. Additionally, we want to explore whether these skews are present in ads for protected categories, including housing, credit, and employment.

Ad setup Previous work showed that, based on the image and the linked website, Facebook steers employment ads towards users whose demographics correlate with the distribution of workers in the market [13]. For example, Ali et al. [13] found that ads for jobs in the lumber industry were delivered disproportionately to white

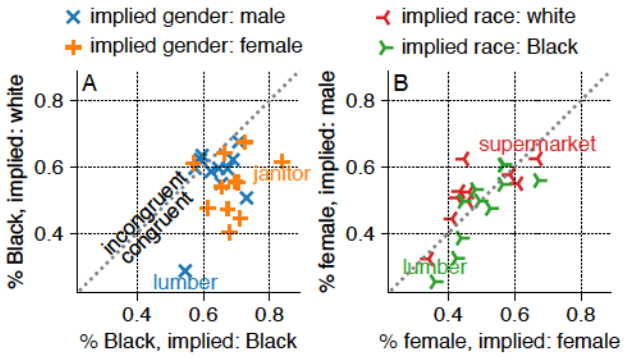

Figure 7: Delivery of ads featuring synthetic faces in various employment ads. Large differences along gender and race lines stem from differences in advertised industries. A) A majority of ads follow a congruent race skew, where images of Black people are more likely to be delivered to Black users. B) We do not see a similar skew for gender.

men. However, their results were inconclusive as to whether the demographics of the person in the ad image could further influence the delivery.

To test this, we obtain stock images—*not* containing an image of a person—related to the 11 job categories as in Ali et. al [13]: AI engineer, doctor, janitor, lawyer, lumber (logger), nurse, preschool teacher, restaurant server, secretary, supermarket clerk, and taxi driver. We super-impose on top of these images the faces generated using StyleGAN 2 in previous experiments, focusing on the adult age images. The experiments we run use only the final images that combine the stock background with a generated face. We advertise each job in four configurations, implying intersection of male and female, white and Black identities. As in previous experiments, we run the ads to a target audience split across Florida and North Carolina, and measure gender and race skew using the marketing API. We provide destination links for each ad as the relevant page on indeed.com, a popular job-hunting website (meaning any user who clicked on our ad would indeed be presented with a page of relevant potential jobs). This is referred to as Campaign 4 in Table 2.

Results Figure 7 shows the results of this experiment. Each tick mark represents a *pair* of ads for the same job but with different identities implied in the image. The x-axis of the left (right) graph is the fraction of delivery to Black (female) users when the image is of a Black (female) person, and the y-axis of the left (right) graph is the fraction of deliver to Black (female) users when the when the image is of a white (male) person. If the demographics of the face in the image did not affect delivery, we would expect that all ads would lie along the dotted $x = y$ line. However, tick marks below the dotted line shows skew in the congruent direction (e.g., images of Black people in employment ads are more likely to be delivered to Black Facebook users, and vice versa), and tick marks above the dotted line show the opposite.

We make a number of observations. *First*, we see that, consistent with prior work [13], the ads in different industries show clear differences in delivery along racial and gender lines. For example, ads in the janitorial industry are delivered disproportionately to Black women. *Second*, we see evidence in Figure 7A that ads show a congruent skew along racial lines. As an example, for jobs in lumber industry delivered to an actual audience that was 55% self-reported Black when the face was of a Black man; the same ad but with the image of a white man delivered to an actual audience that was only 28% Black. We observe that the vast majority of the employment ads delivered with a congruent race skew, though the amount of skew varies by job. *Third*, we do not see as much evidence of the same systematic skews in Figure 7B along gender lines, as the ads are distributed roughly evenly across the $x = y$ line.

Regression Following this informal analysis, we quantify the findings using a series of mixed-effects linear regression models in Table 5. We begin by setting the fraction of the actual audience that is self-reported as Black as the dependent variable and the binary indicator of implied race identity as the independent variable (True for Black, False for white). We build three models: (I) using only the ads that imply a male gender, (II) using only the ads that imply female, and finally (III) by using all the ads. For each, we group the ads by job type to fit separate intercepts (hence the use of a mixed-effects model). The coefficient of the independent variable quantifies the skew, with a positive coefficient indicating a congruent skew.

For example, model (I) shows that a job ad with a picture of a Black woman was delivered to an actual audience whose fraction of Black users was 14.1 percentage point higher than if that same job was advertised with a picture of a a white woman. The effect is significant, yet of lower magnitude, when male gender is implied (model (II)). Models (IV)-(VI) show no statistically significant effects on gender skews, confirming the intuitive findings from Figure 7B.

7 LIMITATIONS

It is important to recognize that while we have uncovered aspects of how perceived demographics affect ad delivery in our experimental campaigns, broad conclusions cannot be made about how this impacts more complex ads, or whether the effects exist (or if so, how large they are) on other platforms. Examples of such ads can be those that include images with a diverse group of faces, races not covered by this study, images of people whose gender presentation does not conform to stereotypes, photos containing other background details, or ads linking to complex content. All of these aspects and others may skew delivery in conjunction with of the perceived identity of the person pictured, leading to potentially different results. Still, the ads we ran were intended to mimic real-world ads, and the skews we observed were clear enough to strongly suggest that similar effects are present for ads run by real-world advertisers. We leave a full exploration of how the demographics of users in ads interact with ads containing a variety of other content, as well as an exploration of the extent to which these effects are present on other ad platforms, to future work.

8 DISCUSSION

The advertising industry has long been perfecting its ability to reach the "right" audience, from ads placed in newspapers read by select demographics and TV ads played at strategically chosen times, to targeting users based on interests inferred from online activity.

	Dep. variable: Fraction Black			Dep. variable: Fraction female		
	Implied: female +	Implied: male ×	overall +, ×	Implied: Black ⋋	Implied: white ⋌	overall ⋋, ⋌
Intercept	0 544	0 570	0 557	0 480	0 489	0 484
Implied: Black	0 141	0 070	0 105	-	-	-
Implied: female	-	-	-	0 023	0 020	0 002
Adj. ^{2}	0 446	0 117	0 288	0 035	0 042	0 024

0 05; 0 01; 0 001

Table 5: Results of mixed-effects regression for modeling the fraction of the actual audience that self-identifies as Black and the fraction of the actual audience that self-identifies as female. We see statistically significant positive coefficients for implied Black images, supporting a congruent race skew, and no significant gender effects.

Due to the growing concerns about potentially harmful effects of *micro-targeting*, online platforms are now removing the most problematic targeting options. However, this does not necessarily mean that ads can no longer be delivered to a very specific group of individuals. Instead, that role is increasingly taken over by the ad delivery algorithms of online advertising platforms. Previous research has shown that delivery optimization can lead to societally negative outcomes such as gender and race skews in employment and housing ads [13] or price discrimination in other contexts [14].

In this work, we focused on the role that demographic information contained *within an ad image* plays in optimizing the ad delivery. We designed a series of experiments that involved using both stock and synthesized pictures of faces that implied various demographic attributes to measure the demographic makeup of actual audiences that Facebook chose to deliver the ads to. We demonstrated that ads which are identical except for a demographic attribute implied in their images are delivered to vastly different actual audiences. Most notably, images of Black individuals are delivered more to Black users; images of children are delivered more to women; and images of younger women are shown disproportionately to older men.

In certain contexts, skewing the delivery of ads towards the individuals whose demographics are represented in the images can work towards overcoming historical inequities. For example, employers seeking to diversify their workforce cannot explicitly target the under-represented demographics. Instead, they may choose to use imagery that suggests who their desired audience may be. We show that, while the kind of job advertised has a major influence on the demographics of the actual audience, implying demographic attributes in the image can still skew the delivery optimization output in the desired direction. The flip-side, however, is that defaulting to pictures of white men may be even more problematic than previously thought. Previous research shows that minority individuals are less responsive to such ads; our current work shows that they are also less likely to even be *exposed* to such ads in the first place. Importantly, a consumer of traditional media can choose to see ads that they are not normally exposed to by buying a newspaper or watching a TV program targeted at a different demographic. In the online services we study, users do not have an easy way of finding out what ads they do not see.

We also observe that delivery of ads involving certain demographics can skew towards entirely different groups. For example, in our experiments pictures of children are predominantly shown to women. Skewing the delivery this way reflects that, historically, women were more likely to engage with such ads than men, but it also reinforces the stereotype of women as caretakers. Notably, images of children are often used in ads that appear "relevant" as they elicit engagement, but in fact exploit and exacerbate health-related anxieties [48]. Further, in our experiments ads with images representing younger women were predominantly shown to older men. Even if that effect accurately represents population-level interests of older men it might be counter to the advertiser's intention and it raises further ethical questions about the limits of optimization. In summary, our findings contribute to the discussion on the interplay between an advertiser's targeting choices and the platform's ad delivery algorithms. We do not speculate about how the effects we observe would be treated under existing anti-discrimination laws, as they can both contribute to and detract from societally-desired outcomes. Nevertheless, our work can bring more in-depth understanding of the delivery algorithms to researchers, advertisers, and society at large so they can make informed decisions about the measurement, design, and consumption of ads.

9 ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers and our shepherd, Umar Iqbal, for their helpful feedback. We would also like to thank Muhammad Ali for running the exploratory experiments and Avijit Ghosh for sharing his expertise on StyleGAN. This work was funded in part by a grant from the NSF grants CNS-1916020 and CNS-1616234, and Mozilla Research Grant 2019H1.

REFERENCES

[1] 12 CFR § 202.4 (b) – Discouragement. https://www.law.cornell.edu/cfr/text/12/202.4.
[2] 24 CFR § 100.75 – Discriminatory Advertisements, Statements And Notices. https://www.law.cornell.edu/cfr/text/24/100.75.
[3] Exhibit A – Programmatic Relief. https://nationalfairhousing.org/wp-content/uploads/2019/03/FINAL-Exhibit-A-3-18.pdf.
[4] Facebook Plans Crackdown On Ad Targeting By Email Without Consent. https://techcrunch.com/2018/03/31/custom-audiences-certification/.
[5] Title VII of the Civil Rights Act of 1964. https://www.law.cornell.edu/wex/title_vii.
[6] 47 USC § 230 – Protection For Private Blocking And Screening Of Offensive Material. https://www.law.cornell.edu/uscode/text/47/230.
[7] About Advertising Objectives. https://www.facebook.com/business/help/517257078367892.
[8] About Audiences For Special Ad Categories. https://www.facebook.com/business/help/2220749868045706.

[9] About bid and budget pacing. https://www.facebook.com/business/help/571961726580148?id=2196356200683573.

[10] About Customer Match. https://support.google.com/adwords/answer/6379332?hl=en.

[11] Laura Alessandretti, Ulf Aslak, and Sune Lehmann. The scales of human mobility. *Nature*, 587(7834):402–407, 2020.

[12] Muhammad Ali, Angelica Goetzen, Alan Mislove, Elissa Redmiles, and Piotr Sapiezynski. All things unequal: Measuring disparity of potentially harmful ads on facebook. In *6th Workshop on Technology and Consumer Protection (ConPro'22). IEEE*, 2022.

[13] 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. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–30, 2019.

[14] Muhammad Ali, Piotr Sapiezynski, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Ad delivery algorithms: The hidden arbiters of political messaging. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 13–21, 2021.

[15] Jisun An and Haewoon Kwak. Gender and racial diversity in commercial brands' advertising images on social media. In *International Conference on Social Informatics*, pages 79–94. Springer, 2019.

[16] Jisun An and Ingmar Weber. Diversity in online advertising: a case study of 69 brands on social media. In *International Conference on Social Informatics*, pages 38–53. Springer, 2018.

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

[18] Derek R Avery, Morela Hernandez, and Michelle R Hebl. Who's watching the race? racial salience in recruitment advertising 1. *Journal of Applied Social Psychology*, 34(1):146–161, 2004.

[19] Sandra L Bem and Daryl J Bem. Does sex-biased job advertising "aid and abet" sex discrimination? 1. *Journal of Applied Social Psychology*, 3(1):6–18, 1973.

[20] Lawrence Bowen and Jill Schmid. Minority presence and portrayal in mainstream magazine advertising: An update. *Journalism & Mass Communication Quarterly*, 74(1):134–146, 1997.

[21] Scott Coltrane and Melinda Messineo. The perpetuation of subtle prejudice: Race and gender imagery in 1990s television advertising. *Sex roles*, 42(5):363–389, 2000.

[22] Custom Audiences From Your Customer List. https://www.facebook.com/business/help/606443329504150.

[23] Amit Datta, Anupam Datta, Jael Makagon, Deirdre K Mulligan, and Michael Carl Tschantz. Discrimination in online advertising: A multidisciplinary inquiry. In *Conference on Fairness, Accountability and Transparency*, pages 20–34. PMLR, 2018.

[24] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on privacy enhancing technologies*, 2015(1):92–112, 2015.

[25] Scott Detrow. What did cambridge analytica do during the 2016 election? *NPR*, 2018. https://www.npr.org/2018/03/20/595338116/what-did-cambridge-analytica-do-during-the-2016-election.

[26] Emily Dreyfuss. Facebook changes its ad tech to stop discrimination. *WIRED*, 2019. https://www.wired.com/story/facebook-advertising-discrimination-settlement/.

[27] Facebook: About The Delivery System: Ad Auctions. https://www.facebook.com/business/help/430291176997542.

[28] Facebook Ad Targeting Options. https://www.facebook.com/business/ads/ad-targeting.

[29] Facebook: Designated Market Areas for Ad Targeting. https://www.facebook.com/business/help/1501907550136620.

[30] Facebook Makes Moves To Stop Discriminatory Ad Targeting. http://www.socialmediatoday.com/social-business/facebook-makes-moves-stop-discriminatory-ad-targeting.

[31] Florida Voter Extract Disk File Layout. https://files.floridados.gov/media/704671/voter-extract-disk-file-layout.pdf.

[32] Danielle Gaucher, Justin Friesen, and Aaron C Kay. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology*, 101(1):109, 2011.

[33] Phillipa Gill, Vijay Erramilli, Augustin Chaintreau, Balachander Krishnamurthy, Konstantina Papagiannaki, and Pablo Rodriguez. Best paper–follow the money: understanding economics of online aggregation and advertising. In *Proceedings of the 2013 Internet Measurement Conference (IMC'13)*, pages 141–148, 2013.

[34] GitHub StyleGAN 2 Repository. https://github.com/NVlabs/stylegan2.

[35] Jessica J Good, Julie A Woodzicka, and Lylan C Wingfield. The effects of gender stereotypic and counter-stereotypic textbook images on science performance. *The Journal of Social Psychology*, 150(2):132–147, 2010.

[36] Good questions, real answers: How does facebook use machine learning to deliver ads? https://www.facebook.com/business/news/good-questions-real-answers-how-does-facebook-use-machine-learning-to-deliver-ads/.

[37] Google: About Audience Targeting. https://support.google.com/google-ads/answer/2497941?hl=en.

[38] Joshua Green and Sasha Issenberg. Inside The Trump Bunker, With Days To Go, 2016. https://www.bloomberg.com/news/articles/2016-10-27/inside-the-trump-bunker-with-12-days-to-go.

[39] The Guardian. Cambridge Analytica: how did it turn clicks into votes? https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie.

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

[41] Scott Highhouse, Sandra L Stierwalt, Peter Bachiochi, Allison E Elder, and Gwenith Fisher. Effects of advertised human resource management practices on attraction of african american applicants. *Personnel Psychology*, 52(2):425–442, 1999.

[42] 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.

[43] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[44] Ava Kofman and Ariana Tobin. Facebook ads can still discriminate against women and older workers, despite a civil rights settlement. *ProPublica*, 2019. https://www.propublica.org/article/facebook-ads-can-still-discriminate-against-women-and-older-workers-despite-a-civil-rights-settlement.

[45] Aleksandra Korolova. Facebook's Illusion Of Control Over Location-related Ad Targeting. Medium, December 2018. https://medium.com/@korolova/facebooks-illusion-of-control-over-location-related-ad-targeting-de7f865aee78.

[46] Linkedin: Account Targeting. https://business.linkedin.com/marketing-solutions/ad-targeting/account-targeting.

[47] Patrick F McKay and Derek R Avery. Warning! diversity recruitment could backfire. *Journal of Management Inquiry*, 14(4):330–336, 2005.

[48] Madhumita Murgia. Time to turn off Facebook's digital fire hose. *Financial Times*, 2021.

[49] Dmitry Nikitko. Learn direction in latent space. https://github.com/Puzer/stylegan-encoder/blob/master/Learn_direction_in_latent_space.ipynb.

[50] North Carolina Laws, Chapter 163: Elections and Election Laws. https://www.ncleg.gov/Laws/GeneralStatuteSections/Chapter163.

[51] North Carolina Voter Extract Disk File Layout. https://s3.amazonaws.com/dl.ncsbe.gov/data/layout_ncvoter.txt.

[52] Panagiotis Papadopoulos, Nicolas Kourtellis, Pablo Rodriguez Rodriguez, and Nikolaos Laoutaris. If you are not paying for it, you are the product: How much do advertisers pay to reach you? In *Proceedings of the 2017 Internet Measurement Conference*, pages 142–156, 2017.

[53] Lesley A Perkins, Kecia M Thomas, and Gail A Taylor. Advertising and recruitment: Marketing to minorities. *Psychology & Marketing*, 17(3):235–255, 2000.

[54] Pinterest: Audience Targeting. https://help.pinterest.com/en/business/article/audience-targeting.

[55] Angelisa C Plane, Elissa M Redmiles, Michelle L Mazurek, and Michael Carl Tschantz. Exploring user perceptions of discrimination in online targeted advertising. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 935–951, 2017.

[56] Filipe N. Ribeiro, Koustuv Saha, Mahmoudreza Babaei, Lucas Henrique, Johnnatan Messias, Fabrício Benevenuto Oana Goga, Krishna P. Gummadi, and Elissa M. Redmiles. On Microtargeting Socially Divisive Ads: A Case Study Of Russia-linked Ad Campaigns On Facebook. In *Conference on Fairness, Accountability, and Transparency*, pages 140–149, Atlanta, Georgia, USA, January 2019. ACM.

[57] Richard Rothstein. *The color of law: A forgotten history of how our government segregated America*. Liveright Publishing, 2017.

[58] Piotr Sapiezynski, Avijit Ghosh, Levi Kaplan, Alan Mislove, and Aaron Rieke. Algorithms that "Don't See Color": Measuring Biases in Lookalike and Special Ad Audiences. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES'22)*, Oxford, United Kingdom, Aug 2022.

[59] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE, 2020.

[60] Shutterstock Stock Photos. https://www.shutterstock.com/.

[61] Craig Silverman. How A Massive Facebook Scam Siphoned Millions Of Dollars From Unsuspecting Boomers. https://www.buzzfeednews.com/article/craigsilverman/facebook-subscription-trap-free-trial-scam-ads-inc.

[62] Craig Silverman and Ryan Mac. Facebook Profits As Users Are Ripped Off By Scam Ads. https://www.buzzfeednews.com/article/craigsilverman/facebook-ad-scams-revenue-china-tiktok-vietnam.

[63] Scott Spencer. An update on our political ads policy, 2019. https://blog.google/technology/ads/update-our-political-ads-policy/.

[64] Thomas H Stevenson. A six-decade study of the portrayal of african americans in business print media: Trailing, mirroring, or shaping social change? *Journal of Current Issues & Research in Advertising*, 29(1):1–14, 2007.

[65] The 2022 Florida Statues, Title IX: Electors and Elections, Chapter 97: Qualification and Registration of Electors. http://www.leg.state.fl.us/statutes/index.cfm?App_mode=Display_Statute&Search_String=&URL=0000-0099/0097/Sections/0097.0585.html.
[66] Giridhari Venkatadri, Elena Lucherini, Piotr Sapiezyński, and Alan Mislove. Investigating sources of PII used in Facebook's targeted advertising. In *Proceedings of the Privacy Enhancing Technologies Symposium (PETS'19)*, Stockholm, Sweden, Jul 2019.
[67] Ellen L. Weintraub. Don't Abolish Political Ads On Social Media. Stop Microtargeting., November 2019. https://www.washingtonpost.com/opinions/2019/11/01/dont-abolish-political-ads-social-media-stop-microtargeting/.

A CONTROLLING FOR ECONOMIC CONFOUNDERS

One concern about our results may be that the effects we describe are not solely due to the race implied in an ad, and that they could potentially be explained by economic factors instead. We first note that there are, indeed, significant population-level economic differences between white and Black people in the U.S, rooted in historical and current systemic racism [57]. Through a legally enforced practice of *redlining*, Black Americans were forced to live in less attractive areas and barred from obtaining mortgages and, as a result, did not have the same opportunities to accumulate wealth through home ownership. Despite the introduction of the Fair Housing Act of 1968, the Equal Credit Opportunity Act of 1974, and the Community Reinvestment Act of 1977, Black Americans continue facing discrimination even in legally protected areas. Because of the resulting inequality, even if a phenomenon is "color-blind" and based on economic status, it might affect people of different races in a disparate way.

Regardless, we attempted to construct an experiment in which we (partially) control for the economic status to better tease out the racial effect in ad delivery. We do not have information on the individual level economic status. Instead, for each individual we use poverty rate of the ZIP code of their residence as a proxy. In the audiences we targeted originally, half of the white people we targeted lived in ZIP code with poverty at 12% or below, and half of the Black people lived in ZIP codes with poverty at 16% or below, and the difference between mean ZIP code-level poverty was statistically significant. We then subsampled the audiences such that there was an identical distribution of ZIP code-level poverty between all the intersectional race gender state groups. The new audience had 1,730,212 individuals from each state, down from 2,870,772, and maintained the intersectional stratification of gender, race, and age described in the paper. We ran our suite of 100 stock images for North Carolina Black voters/Florida white voters as the first campaign and North Carolina white voters/Florida Black voters as the second campaign, in the same process as prior experiments. However, Facebook rejected over 95% of the ads, and after repealing their decision still rejected 44 of the ads from either campaign. This is despite all 100 of these ads being run previously and there being no differences between different ads other than their images; in fact, many of the images rejected in one campaign ran at the same time in another. While we do not know the reason our ads were rejected, due to time constraints we proceeded with the analysis. We removed all 44 advertisements rejected from either campaign from both campaigns, and further subsampled the ads so that the age and gender of the ads were not correlated with race. This left us with a total of 24 ads from each campaign. Our results for the poverty-controlled experiment exhibited similar trends to the results we found in the main experiments; whether the image is of a Black person is statistically significant in predicting the delivery to Black users, see Table A1.

It is important to note that this experiment cannot be directly compared to our previous experiment which was not poverty-controlled. The two experiments were run at different times and Facebook rejected 44 ads in the poverty-controlled experiment which were all accepted previously, resulting in a smaller sample of ads for analysis. Our results here suggest similar trends even when controlling for poverty level, but due to the constraints mentioned, we cannot decide conclusively to what degree controlling for poverty explains our trends. Nonetheless, we find that due to the relationship between poverty and race in the United States our findings hold merit and strongly suggest a relationship between the race of the person in an ad and the delivery statistics of that ad.

	% Black
Intercept	0.6171
Black	0.0849
Female	0.0186
Teen	0.0111
Middle-aged	0.0388
Elderly	0.0066
R^2	0.392

$p < 0.05$; $p < 0.01$; $p < 0.001$

Table A1: Linear regression results for stock photos delivered to Facebook users while controling for economic differences between intersections of gender and race.