Title: Race Discrimination in Internet Advertising: Evidence From a Field Experiment

Authors: Neil KR Sehgal[1], Dan Svirsky[2*]

Affiliations:
[1] Institute for Applied Computational Science, Harvard University; Boston, MA, USA.
[2] Harvard Business School; Boston, MA, USA. Current affiliation: Uber Technologies, Inc.; Boston, MA, USA. This project arose out of research that began before my employment with Uber Technologies. Uber Technologies did not oversee or financially support this project, and the views and opinions in this paper do not necessarily reflect the views and opinions of Uber Technologies.

*Corresponding author. Email: dansvirsky@gmail.com

Keywords: Racial bias, Discrimination, Facebook, Advertising

We present the results of an experiment documenting racial bias on Facebook's Advertising Platform. We find that darker skin complexions are penalized, leading to real economic consequences. For every $1,000 an advertiser spends on ads with models with light-skin complexions, that advertiser would have to spend $1,159 to achieve the same level of engagement using photos of darker skin complexion models. Facebook's budget optimization tool reinforces these viewer biases. When pictures of models with light and dark complexions are allocated a shared budget, Facebook funnels roughly 64% of the budget towards photos featuring lighter skin complexions.

Main Text:

Many important marketplaces that used to operate in physical spaces have moved online, but the problem of racism persists. Ads shift from billboards to timelines. Wedding photographers advertise on Instagram instead of the Yellow Pages. These changes are economically important: half of the $225 billion spent on U.S. advertising in 2019 was spent on just three online platforms: Amazon, Google, and Facebook *(1)*.

As these markets have moved online, research documenting race discrimination has continued to find that racial biases have an impact on important outcomes, online and off. Discrimination by ethnicity has been well-documented in marketplaces from the market for credit, goods, labor, short-term rentals, crime enforcement, and housing *(2-8)*. The growth towards online markets poses new legal challenges in a country where antidiscrimination law was designed before the internet was developed.

This paper presents an experiment documenting the role of race discrimination on Facebook's advertising platform. We measure whether photographs of people with darker complexions garner less engagement -- defined as likes divided by audience views -- and whether, as a result, are more costly to run. We further measure whether Facebook's optimization algorithms contribute to discrimination or ameliorate it. In the experiment, we run advertisements for wedding photographers using photographs of models that vary in their skin complexion. We use a 2x2 design: we compare pairs of photographs that are similar in every way except skin complexion to get a baseline measure of differences in engagement. Then we run advertisements using photographs from the same pairing, but zoomed in and cropped so as to remove non-race related features (e.g., details of a dress or a flower arrangement) and make the skin complexion of the models a more salient part of the ad. We find that when advertisements highlight subjects with darker skin, they receive 10.39% fewer likes. This difference has economic significance: advertisers must spend 11.59% more per photo to garner the same engagement for a picture highlighting a person of color. Observational data on the demographics of the audience suggest that Facebook is actively showing the different ads to different types of users, but we do not find evidence that these under-the-hood decisions make the treatment effect stronger.

We also find that Facebook's budget optimization tool, while on its face neutral, exacerbates discrimination by reflecting user bias. Facebook offers advertisers a tool to optimize budget decisions by spending more money on ads that get more engagement. We find that when this tool is turned off, the advertising budget is spent equally across all ads, regardless of the model's skin complexion. But when the tool is turned on, Facebook funnels roughly 64% of advertising dollars towards pictures of models with light skin complexions, a nearly 2:1 advantage.

Our findings have implications for the legal regulation of online markets, because they show how facially neutral algorithms can reinforce user biases to make racial disparities worse. Our findings also contribute to the social science literature on discrimination by showing that racial disparities exist even in settings where statistical discrimination is a less natural explanation, compared to straightforward taste-based animus. Finally, we use Facebook's data on the location of the user to assess whether our measure of discrimination correlates across geographies with other measures of discrimination, like the Implicit Association Test.

This paper is organized as follows. Section II describes the experimental design. Section III presents results. Section IV concludes.

Experimental Design

We conduct an experiment to measure whether ads for wedding photographers featuring people with darker complexions get fewer likes than ads for wedding photographers featuring people with lighter complexions. We further test whether Facebook's optimization tools affect any underlying disparities we find. We also measure whether, as a result of any disparity, photos featuring people with darker complexions require higher advertising costs to garner similar levels of audience engagement. We pre-registered the experiment on osf.io, the Center for Open Science's repository. All code and data is available in this repository as well.

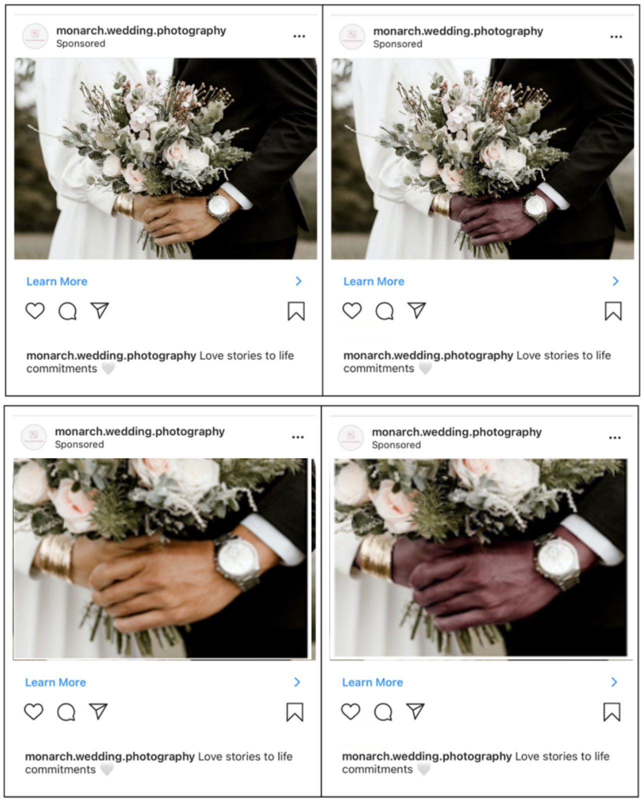
1. *Image Selection and 2x2 Design*

Ideally, we would take two photos that are identical in everything except the skin color of the subject and measure whether changing the skin color leads to fewer likes. One could compare ads that have models who look similar, in a similar pose and context, but who vary by complexion. Or, one could take an ad with someone with light skin and make the skin complexion look darker, or someone with dark skin and make the skin complexion look lighter, using photo editing software. Then the research question is straightforward -- do two otherwise similar pictures have different engagement levels when the skin color changes?

But such an approach is imperfect because it is impossible to only change skin color without affecting other attributes of the photograph. Shadows, lighting, contrast, hues, saturation -- all of these traits are important to the aesthetic value of a photograph and are hard to control or measure by the experimenter. Pictures, after all, are worth an aphoristic thousand words. And using any one pair of pictures raises external validity concerns -- would such findings apply to other pictures?

We address this challenge using the 2x2 experimental design illustrated below. Consider the two pictures in the left column -- a zoomed out picture of a bride and groom embracing. Simply measuring the difference in Likes for the darker skin versus the lighter skin suffers from the problem described above: unobserved or unmeasured differences between the pictures that correlate with skin tone.

Our measure of racial discrimination is slightly different. Given a baseline difference in Likes for two similar pictures with different skin complexions, we ask whether there is a racial penalty for zoomed-in versions of these same photographs, where the skin takes up a bigger portion of the picture and is therefore more salient, *relative to the baseline racial difference of the zoomed-out versions of these same photographs.*



monarch.wedding.photography
Sponsored
Learn More
monarch.wedding.photography Love stories to life commitments

monarch.wedding.photography
Sponsored
Learn More
monarch.wedding.photography Love stories to life commitments

monarch.wedding.photography
Sponsored
Learn More
monarch.wedding.photography Love stories to life commitments

monarch.wedding.photography
Sponsored
Learn More
monarch.wedding.photography Love stories to life commitments

Fig. 1. Illustration of the 2x2 experimental design. In the experiment, we compare two similar pictures where skin complexion differs to measure differences in how many users "Like" each ad on Instagram. We conduct a 2x2 design by then zooming in on each picture in a way that makes the skin more salient and then measuring whether this creates any disparities in "Like" rates between the two pictures, after controlling for any baseline differences in "Like" rates.

The dependent variable of interest is the difference in the proportion of likes for the two ads, when we present a zoomed-in photograph where skin complexion is more salient, relative to the difference between the baseline (zoomed-out) pictures.

Specifically, let P_L be equal to the number of likes for a photograph of a model with a lighter complexion divided by the total number of people who viewed that photograph, P_D be the same for a photograph of a model with a dark complexion, P_LZ be the same for the zoomed-in photograph of the model with a light complexion, and P_DZ be the same for a zoomed-in photograph of the model with a dark complexion. The change in {Likes / Views} when an ad zooms in and makes the skin in a picture more salient can be expressed, for people with light skin, as follows:

$$(P_{LZ} - P_{L}) \tag{1}$$

And similarly for people with darker complexions.

We are most interested in whether the following equation holds:

$$(P_{LZ} - P_{L}) - (P_{DZ} - P_{D}) = 0 \qquad (2)$$

If equation (1) does not hold, and if the expression on the left is positive, then this is evidence that making the skin more salient penalizes ads with people with darker complexions. If the expression on the left is negative, then this is evidence that making the skin more salient penalizes ads with people with lighter complexions. If the expression holds, then this is evidence of a null effect.

Another reasonable measure of discrimination is in equation (3):

$$(P_{LZ} + P_{L}) - (P_{DZ} + P_{D}) = 0 \qquad (3)$$

This equation tests whether the total engagement rate for photos with models with light complexions is equal to the total engagement rate for photos with models with dark complexions. If the expression is greater than zero, that suggests a bonus when the photo features lighter complexions.

We consider Equation (2) a cleaner test, as it measures whether darker complexions see a penalty when skin is more salient relative to a baseline that tests underlying differences in the pictures themselves. For that reason, we pre-registered Equation (2) as the primary outcome of interest. Nonetheless, Equation (3) may be of independent interest, and we report both.

As noted above, the best approach for this experiment is not obvious ex ante -- similar photographs of models with different complexions? An identical photograph with the complexion changed in Adobe Photoshop? We address this challenge with an all-of-the-above approach. We run the experiment six times, with six pairs of photographs. In Pairs 1 and 2, the photographs are of different subjects but holding the same pose. In Pairs 3 and 4, the photos are identical, but the skin complexion in an original photo was made darker with Adobe Photoshop. In Pairs 5 and 6, the photos are identical, but the skin complexion in an original photo was made lighter with Adobe Photoshop.

Using six sets of photos also helps to address the second challenge described above -- external validity. If we find a penalty for skin complexion for one pair of photographs, but not for another, then this is not clear evidence of discrimination.

Figure 2 summarizes the six tests we run.

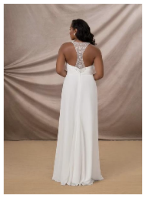
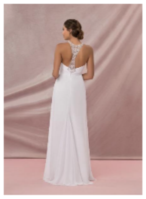
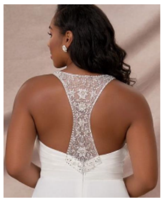
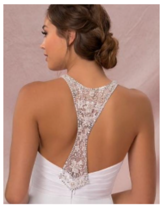
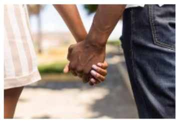
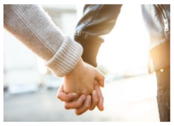
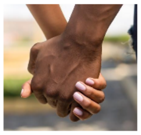
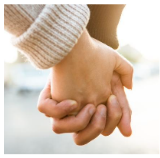
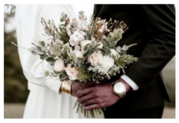
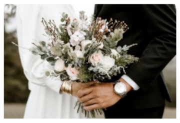
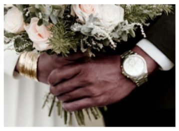
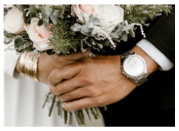
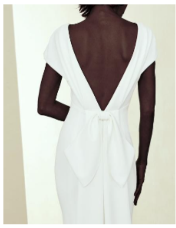
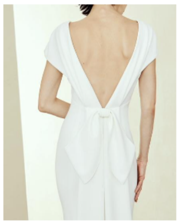
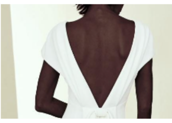
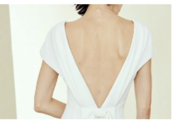
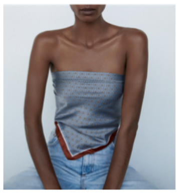
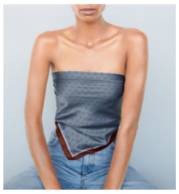
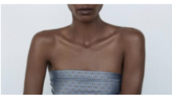
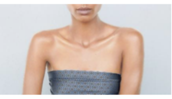
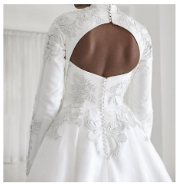
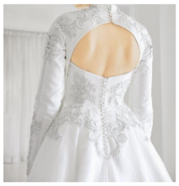
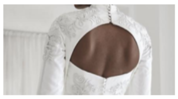
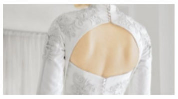
	Pair 1		Pair 2	
Modification	**Dark Complexion, uncropped**	**Light Complexion, uncropped**	**Dark Complexion, cropped**	**Light Complexion, cropped**
No artificial lightening or darkening				
				
Light skin artificially darkened				
				
Dark skin artificially lightened				
				

Fig. 2: Images tested[1]. We test six sets of images in the experiment, running four ads for each of the six sets. Each set has four pictures: two similar pictures where the skin complexion differs, then two identical pictures, but zoomed in. In two of the sets, we find pictures that look similar but with different models with different skin complexions. In two of the sets, we take one picture and use Adobe Photoshop to artificially make the skin complexion look lighter. In the remaining two sets, we take one picture and use Adobe Photoshop to artificially make the skin complexion look darker.

[1] Images available under Creative Common License BY-NC-ND 4.0.

2. *Facebook Advertising Platform*

Facebook is one of the dominant online advertising platforms, with nearly $70 billion in advertising revenue in 2019 *(9)*. It is the second-largest digital ads platform, behind Google. Along with Facebook, these three platforms account for 90% of digital ads and more than half of all U.S. ad spending *(1)*.

Facebook exercises significant control over who sees each advertisement in ways that are opaque. In theory, an advertiser could simply give Facebook a list of phone numbers and ask the platform to randomly select a subset of this audience to view the ad. But recent research suggests Facebook does not always do this. Ali et al (2019) show that Facebook uses its user data -- a person's likes, friends, interests, and so on -- to target ads based on the content of those ads, effectively choosing the audience reach based on who it thinks will respond to the ad *(10)*.

In addition, Facebook lets advertisers select their own criteria for targeting ads, albeit with some limitations following a civil rights lawsuit. An advertiser can aim their ad at age groups, by geography, by interests (like sports or wedding photography), or by whether a person is similar to an existing group of Facebook users. Facebook determines user interests through multiple factors including past page and advertisement engagement, demographics, and network connection speed.

When creating an ad, Facebook offers 11 unique Campaign Objectives for users to choose. Based on the selected objective, Facebook serves ads to different audiences based on who it believes is most likely to take a desired action. Examples of objectives include audience reach, which tries to show the ad to as many people as possible, and audience engagement, which tries to maximize the number of Likes, comments, and shares for an ad.

3. *Advertisement creation*

We run an ad on Instagram for each of the twenty-four images, using the Facebook Ad Manager interface. We focus on the Engagement objective, which targets users most likely to engage with an ad through follows, comments, shares, or likes. We set the audience to Instagram users in the United States, age 18 years or older, and with an interest in wedding photography. This specific audience yielded an eligible audience of 9.6 million Instagram users. For comparison, an audience of Instagram users in the United States, age 18 years or older, but with no specified interests yielded an eligible audience of 130 million Instagram users.

We place ads exclusively on the Instagram Feed. Advertisements in the Instagram Feed are identical to normal posts, except for a small "Sponsored" disclosure and a clickable "Learn More" link, directing viewers towards our ad account's profile page. Within each group, the four ads possess an identical account name and caption, only differing by image. Given the large eligible audience of millions of users, it is unlikely any individual Instagram user would have seen more than a single ad from one grouping. Figure 1 shows an example of an advertisement set. Based on data from pilot testing for our selected audience, an ad is viewed by 1000 Instagram users for every $15.86 spent. We budgeted each ad to be shown to 1160 users ($18.39) over 24 hours based on a power calculation with Beta set to 0.2 and Alpha set to 0.05 for an effect of a five percentage point difference in likes between two pictures (also based on pilot testing).

Results

1. *Main Experiment Results: Who Saw the Ads*

In total, 34,419 Instagram users viewed one of the 24 advertisements. Users responded with a Like in 7,530 cases for an average like/ad view of 0.22. Summary statistics for the demographics of the advertisement viewers are listed in Table 1. A majority of the viewer population is female and age 18-24.

		Ad Views			
		Dark Complexion		**Light Complexion**	
	All Ads	**Cropped**	**Uncropped**	**Cropped**	**Uncropped**
Gender					
Female	23,462 (68.2%)	3,776 (52.5%)	8,383 (80.3%)	2,482 (39.2%)	8,821 (84.4%)
Male	10,697 (31.1%)	3,356 (46.7%)	1,987 (19%)	3,796 (59.9%)	1,558 (14.9%)
Unknown	260 (0.8%)	57 (0.8%)	76 (0.7%)	57 (0.9%)	70 (0.7%)
Age					
13-17	1 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0%)
18-24	17,222 (50.04%)	2,955 (41.1%)	5,701 (54.6%)	2,099 (33.1%)	6,467 (61.9%)
25-34	8,236 (23.93%)	1,791 (24.9%)	2,657 (25.4%)	1,479 (23.3%)	2,309 (22.1%)
35-44	3,925 (11.4%)	1,094 (15.2%)	965 (9.2%)	1,125 (17.8%)	741 (7.1%)
45-54	2,603 (7.56%)	706 (9.8%)	562 (5.4%)	881 (13.9%)	454 (4.3%)
55-64	1,606 (4.67%)	439 (6.1%)	360 (3.4%)	519 (8.2%)	288 (2.8%)
65+	826 (2.4%)	204 (2.8%)	201 (1.9%)	232 (3.7%)	189 (1.8%)

Table 1. Breakdown of Characteristics by Treatment Status. This table shows the demographics of ad viewers, both across all 24 advertising campaigns and within each of the four types of advertisements (cropped, dark complexion; uncropped, dark complexion; cropped, light complexion; and uncropped, light complexion).

The most striking result is that Facebook's advertising platform is making different under-the-hood decisions about which groups to serve each ad to. This is not a novel finding. Past research has documented that even when Facebook's Advertising Platform is given a list of randomly chosen American phone numbers and all ad targeting is turned *off*, Facebook still directs, for example, makeup ads to women *(10)*. Table 1 replicates this finding. It is not publicly known how Facebook's algorithms make these choices. For purposes of this paper, it means that any treatment effect we find -- any racial bias -- could be driven by user choices, by Facebook's decision about who to serve ads to, or by a combination of the two. Section III.3 delves into this topic.

2. *Main Experiment Results: Treatment Effect*

We find evidence of a penalty for pictures of models with darker complexions. Table 2 presents the results of a regression measuring the treatment effect. Column 1 is a linear regression regressing whether the ad viewer Liked the picture on three variables: whether the picture has a model with darker skin complexion, whether the picture is cropped, and the interaction of these two. The variable of interest is the interaction between the complexion featured in the picture and whether the picture is cropped. Table S1 shows these results in more detail, with exact Likes and Ad Views displayed for each set of ads.

The baseline photos on average have similar amounts of engagement, garnering almost identical levels of likes per view. This is important because if the baseline pictures were wildly different in terms of engagement, it would make interpretation of any results more challenging.[2] But when the ads zoom in, making skin more salient, the photos of people with lighter complexions receive a significantly higher boost relative to the photos of people with dark complexions. The boost is roughly twice as large: the engagement rates jump from 18.4% to 30.8% when the photos feature people with light complexions, as opposed to a jump of 18.3% to 24.1%. Column 1 of Table 2 illustrates these results in a linear regression.

This finding is robust across all six groups of the pictures tested. Table S2 shows the same results, disaggregated by photo. In the baseline (zoomed out, uncropped) images, we see no statistically significant difference in 3 of the 6 sets, a statistically significant bias towards darker complexions in 2 of the 6 sets, and a statistically significant bias towards the lighter complexion in the remaining 1 set. But in all six cases, when the ads zoom in, the picture with models with darker complexions is penalized more (or improves less), relative to its companion picture with models with lighter complexions.

[2] For example, if the baseline light complexion picture saw a 10% engagement rate, and the dark complexion picture saw a 90% engagement rate, then any differences from baseline would be hard to interpret

3. *Is Facebook Causing Discrimination?*

The treatment effect could be explained by discriminatory users, by Facebook's decision about who sees the ads, or some combination of the two. As noted above, Facebook is actively choosing to serve the ads to different populations who differ along observable demographics of gender and age (other demographics, such as race, are not shared with advertisers). This could help explain the disparities we find just as much as user behavior.

We can test this more accurately -- albeit imperfectly -- in two ways: by measuring the treatment effect after controlling for user demographics, and by assessing the treatment effect in an exploratory round of this experiment when Facebook turned its audience optimization off.

First, we test whether the treatment effect persists after controlling for user demographics. If the treatment effect dissipates when controlling for observable demographics, this suggests it is Facebook's choices that are driving the racial disparity. If the treatment effect is stable, then this suggests either that user behavior, not Facebook audience decisions, are driving the disparity, *or* that Facebook's decisions are driving the disparity but in a way that we cannot observe.

Table 2 presents the results of a regression measuring the treatment effect, when we do and do not control for observable demographics. On univariate analysis, likes were associated with a number of image characteristics and demographic categories. On multivariate analysis (Table 2), a higher proportion of likes are independently associated with cropped image, whether the viewer is female, and the viewer's age category. The treatment effect is the interaction between the complexion featured in the picture and whether the picture is cropped.

	Model 1	Model 2
Intercept	0.185 (0.004)	0.193 (0.010)
Is Darker Complexion	-0.001 (0.006)	-0.016 (0.014)
Is Cropped	0.122 (0.007)	0.098 (0.007)
User is Female	--	-0.032 (0.008)
User Age Category (1 - 6)	--	0.011 (0.003)
Is Darker Complexion * Is Cropped	-0.064 (0.009)	-0.056 (0.010)
Is Darker Complexion * User is Female	--	0.001 (0.010)
Is Darker Complexion * User	--	0.006 (0.004)

Age Category		
Adj R^2	0.013	0.017
N	34,159	34,159

Table 2: Linear regression of engagement rate on photo and viewer characteristics. Standard errors clustered at the (picture complexion * cropped) level. The age variable is a categorical variable ranging from 1 to 6, with 1 being ages 18 - 24, 2 being ages 25-34, 3 being ages 35-44, 4 being ages 45-54, 5 being ages 55-64, and 6 being ages 65+.

Table 2 shows that the main treatment effect holds, even when controlling for what we know about the audience demographics. Without controlling for audience demographics, we find a treatment effect of roughly 6.4 percentage points. When controlling for user gender and age category, the treatment effect diminishes slightly, to 5.6 percentage points, but is not statistically significantly different from the treatment effect when not controlling for audience demographics.

Second, we can see whether the treatment effect persists when Facebook turns its audience optimization algorithms off. As described in the supplementary materials, we ran a version of this experiment across all 50 states to assess the geographic variation in the treatment effect. When we conducted this experiment, an unexpected outcome was that Facebook sent a warning message that it would not be able to run its audience optimization algorithm because this advertising campaign was running so many simultaneous ads. Facebook states that when an advertiser runs too many ads at once, 1,200 in our case, ads are unable to be optimized properly and can deliver less often with worse results. Facebook recommended that for an account of our size, the upper limit for good performance is 250 ads.

Because this was unexpected, we did not pre-register this test, so it should be considered exploratory. Nonetheless, we can see how ad engagement rates -- and the treatment effect -- changed during these tests.

The average Like/Ad View is significantly lower in the geographic experiment (0.12) when compared to the main experiment (0.22), which suggests that Facebook's audience optimization tools are effective. However, the main treatment effect persists, with the light complexion pictures improving by 12% when cropped, while the dark complexion pictures see a slight penalty of 0.8% when cropped, as shown in Table S3.

There are important limitations to both approaches. The first test, which shows that the treatment effect persists when controlling for demographics, is limited because there are many user traits we do not observe, especially ethnicity, but also traits such as socioeconomic status and political attitudes. The second test, which shows that the treatment effect persists when Facebook turns its audience optimization tool off and the ads are run in 50 states, is also limited because we have little information on what, exactly, Facebook is doing when it sends this warning. Other optimization tools could still be on, for example.

In sum, further testing helps tease out the role of user-level discrimination and Facebook market design decisions on disparities. It provides provisional evidence that racial disparities are either driven by user preferences or by a combination of user preferences and Facebook decisions that are not observable to the experimenters.

4. Facebook Budget Optimization Tool Leads to Racial Disparities in Ad Spending

Advertisers can explicitly ask Facebook for help in choosing how to spend their advertising budgets. The tool works as follows. Consider an advertiser who has two advertisements to show during a campaign. Suppose the advertiser does not know if one ad works better than the other. Facebook helps the advertiser optimize her budget choices. First, both ads would be displayed to audiences, but Facebook can learn based on user engagement if one ad is more effective and then funnel money towards that advertisement. We take advantage of this to test whether Facebook's optimization tool leads to different spending levels for photos that feature people with different skin complexions when the feature is ON versus when it is OFF. Importantly, this is an exploratory data analysis, since this test was not pre-registered and therefore not part of our main experiment.

When we do this, we find that Facebook automatically funnels the advertising budget towards pictures of people with light skin complexions, presumably to maximize audience engagement. Figure 5 shows results. When budget optimization is turned off, Facebook allocates the total budget identically across the four conditions (light complexion and uncropped, light complexion and cropped, dark complexion and uncropped, and dark complexion and cropped). Each condition receives roughly 25% of the entire advertising budget. As a result, $124.52 is spent on photos of models with light complexions, as compared to $124.40 on photos of models with dark complexions. But when optimization is turned on, Facebook automatically funnels money towards the photos of models with lighter skin complexions, which receive nearly two-thirds of the entire budget instead of half. $159.43 is spent on photos of models with light complexions versus $89.70 on photos of models with dark complexions.[3] Hence, disparities in outcomes driven by user preference can get amplified by Facebook's budget optimization tools.[4] Table S5 shows the same results, with Likes and Ad Views displayed.

[3] While we do see photos of light skin models outperform dark skin models for both cropped and uncropped images, it is not clear why Facebook funnels more money to uncropped images which receive lower likes/view.

[4] This finding is distinct from work by Lambrecht and Tucker (2016) which finds that even when an ad is designed to be shown in a gender-neutral way, cost optimization algorithms show the ad to more men because the male audience was less desirable (and therefore cheaper to target) *(13)*.

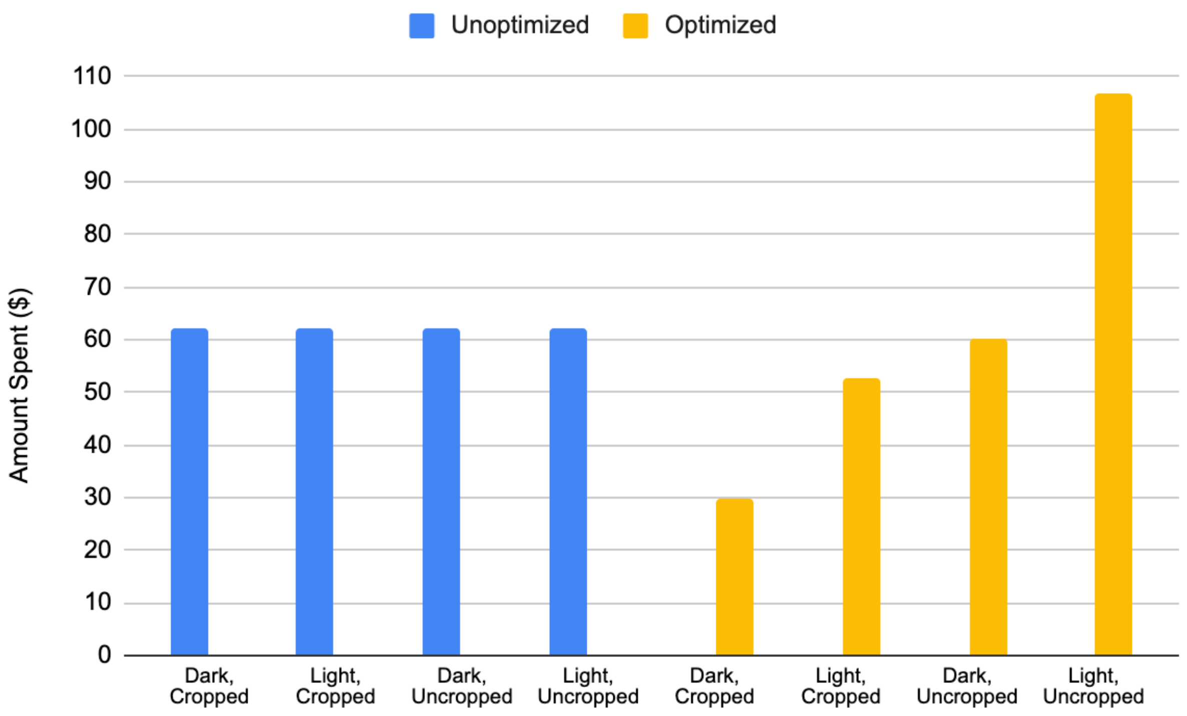


Fig. 3. Optimization Results: This Figure measures how advertising budget is distributed across different pictures when Facebook's budget optimization feature is turned on or off. P-value < 0.001 for Chi-Squared test comparing Amount Spent in first four bars (Optimization OFF) versus Amount Spent in last four bars (Optimization ON).

5. *Estimating the Economic Cost of the Racial Penalty*

The difference in Like rates translates to economic penalties for advertisers who use models with dark complexions. Facebook's Ad Manager lets us calculate this cost more precisely. The basic question we assess is -- if one advertiser spends $1000 promoting a photo of someone with a light complexion to get some level of engagement, how much more would she have to spend to get the same level of engagement if the photo highlighted a model with a dark complexion?

To fix ideas, consider an advertiser who wants to spend some amount on an ad with a model with a light complexion, C_W. That cost, C_W, will allow the advertiser to obtain some number of Likes, L_W. Facebook, as the designer of the advertising marketplace, can define exactly how much money it will cost for advertisement i to reach some level of viewers, R_i. Our key assumption is that since Facebook has control over this because it controls what viewers see, it can assign a fixed R_i for all i. We can then use the experiment results to calculate the Likes per Reach -- L_W / R_W. Combining all these lets us translate C_W -- the amount an advertiser spends -- into L_W -- the amount of engagement the advertiser gets.

Pictures of models with light complexions received Likes 3,880 times out of 16,784 Reaches (23.1% of the time). The pictures of models with dark complexions received Likes 3,650 times out of 17,635 (20.7% of the time). Hence, for every $1,000 that an advertiser spends to achieve a fixed level of engagement for pictures with models with light complexions, an advertiser using pictures of models with dark complexions would have to spend 11.59% more, or $1,159, to achieve the same result.

Another way to measure the size of the racial penalty is by measuring how much Facebook's optimization algorithms reward pictures of people with lighter complexions. As described above, Facebook's budget optimization tool funnels roughly two-thirds of the total advertising budget to pictures of people with light skin complexions. When the tool is turned off, the advertising budget is spent equally. This alternative measure suggests that when a picture is changed so that it highlights a model with lighter complexion, Facebook's optimization tool will penalize it in terms of number of viewers, garnering an audience that is half the size of what it would be if the model had lighter skin.

There are limits in how to interpret this penalty. The effectiveness of advertising is very much a complicated and open question in the economics literature. Hence, one could imagine a picture with models with darker skin that garners fewer Likes, but does better on some other metrics, like conversions. Perhaps fewer people click "Like", but are more likely to then visit the wedding photographer's page, send a message, and hire her. Hence, our calculation here needs to be interpreted with caution, as it only speaks to how much money an advertiser would have to spend to get a level of engagement defined in a specific way.

Discussion

This paper presents the results of an experiment measuring racial bias in a dominant online advertising market. Given two identical or nearly-identical photographs that vary by the model's skin complexion, making the skin more salient by cropping the picture leads to a 21.75% penalty when the model's skin is darker. This directly translates to higher costs for advertisers who feature people with dark skin complexions. We estimate that the racial penalty is associated with a rise in advertising costs of 11.59%.

Since Becker (1957), an important question in the literature on the economics of discrimination is about the mechanism causing disparities -- do people discriminate because of animus *(11)*? Because of statistical inferences about how race correlates with other traits? Both? This experiment adds to that research by finding a racial penalty in a context where taste-based discrimination is a much more natural explanation than statistical inference. In most of the economics literature, racial penalties are found in contexts where the person discriminating might well be making a (perhaps errant) statistical inference using race as a proxy for some other outcome *(12)*. Here, the discrimination we document is simple and instantaneous -- tapping a “Like” button in response to a photograph. If tapping “Like” signifies nothing more than an aesthetic response to a picture, then this is a clean measure of taste-based discrimination. If users tap “Like” for other reasons -- such as to manipulate the types of advertisements that Facebook shows -- then the disparities we find would be driven by something more complicated.

This paper also contributes to the literature seeking a novel method to measure racial animus at scale. This is an increasing challenge because racism is, more and more, too unacceptable to admit to publicly. Across geographies, our measure of a racial penalty can be measured at scale and shows little correlation with existing measures, such as self-reported racial attitudes.

Finally, the experiment also highlights new challenges that regulators and marketplace designers face in addressing a very old problem. Has Facebook done anything illegal? The answer is far from clear. In one sense, Facebook is a market-place designer offering an effective way to reach audiences. Facebook plays a large role in how to serve these ads, trying to optimize for engagement and helping advertisers optimize their budgets. We find strong evidence that Facebook manipulates which audiences see the ad, and that this choice of audience differs based on skin complexion. But we do not find clear evidence that this improves or worsens engagement rates by skin complexion. We find stronger evidence that Facebook’s budget optimization tool does, indeed, exacerbate discrimination, by funneling advertiser money towards photos of models with lighter complexions.

Limits to the experiment suggest further avenues of research. Past research has documented how colorism and racism both operate to harm African Americans in the criminal justice system (Wickett 2021) *(15)*. This experiment does not disentangle the two. It most directly tests the impact of colorism by manipulating skin tone, but because skin complexion is a proxy for racial groups in the United States, our findings could also be driven by racism, either against African Americans or other groups with darker skin tones. This is a common limitation in audit studies, which always focus on manipulating a feature like name or physical appearance in a way that proxies for race. Future research could explore this, even with the same experimental design. In addition, as with other audit studies, this experiment captures a harm whose longer-term implications are unclear. If African-American applicants don’t get callbacks for job interviews, does this translate to lower wages? If models with darker skin complexions are penalized in the engagement with their ads, does this lead to less work? Understanding these dynamics is important, but outside the scope of this paper.

In sum, this paper explores how an old beast evolves when its market environment changes. User-level discriminatory attitudes are nothing new, nor is targeted advertising. What is new is the way that facially neutral algorithmic decision-making can mix with user-level discrimination to pose new legal and ethical challenges.

References and Notes

1. K. H. and S. Vranica, How Covid-19 Supercharged the Advertising “Triopoly” of Google, Facebook and Amazon. *Wall Street Journal* (2021), (available at https://www.wsj.com/articles/how-covid-19-supercharged-the-advertising-triopoly-of-google-facebook-and-amazon-11616163738).

2. D. G. Pope, J. R. Sydnor, What’s in a Picture? *Journal of Human Resources*. **46**, 53–92 (2010).

3. J. L. Doleac, L. C. D. Stein, The Visible Hand: Race and Online Market Outcomes. *The Economic Journal*. **123**, F469–F492 (2013).

4. D. Pager, The Mark of a Criminal Record. *American Journal of Sociology*. **108**, 937–975 (2003).

5. A. Agan, S. Starr, Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment*. *The Quarterly Journal of Economics*. **133**, 191–235 (2017).

6. B. Edelman, M. Luca, D. Svirsky, Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. *American Economic Journal: Applied Economics*. **9**, 1–22 (2017).

7. W. C. Horrace, S. M. Rohlin, How Dark Is Dark? Bright Lights, Big City, Racial Profiling. *Review of Economics and Statistics*. **98**, 226–232 (2016).

8. A. Hanson, Z. Hawley, Do landlords discriminate in the rental housing market? Evidence from an internet field experiment in US cities. *Journal of Urban Economics*. **70**, 99–114 (2011).

9. Facebook, Facebook Reports Fourth Quarter and Full Year 2019 Results. Facebook Investor Relations. (2020) https://investor.fb.com/investor-news/press-release-details/2020/Facebook-Reports-Fourth-Quarter-and-Full-Year-2019-Results/default.aspx.

10. M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, A. Rieke, Discrimination through Optimization. *Proceedings of the ACM on Human-Computer Interaction*. **3**, 1–30 (2019).

11. G. S. Becker, *The Economics of Discrimination* (The University Of Chicago Press, Chicago ; London, 1957).

12. J. A. Bohren, A. Imas, M. Rosenberg, The Dynamics of Discrimination: Theory and Evidence. American Economic Review. **109**, 3395–3436 (2019).

13. A. Lambrecht, C. Tucker, Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science*. **65**, 2966–2981 (2019).

14. R. Block Jr., C. Crabtree, J. B. Holbein, J. Q. Monon, Are Americans less likely to reply to emails from Black people relative to White people?. *Proceedings of the National Academy of Sciences.* **118** (52): e2110347118 (2021).

15. A. Wickett. 2021. Not so Black and White: An Algorithmic Approach to Detecting Colorism in Criminal Sentencing. *ACM SIGCAS Conference on Computing and Sustainable Societies*, **COMPASS '21**, 46 (2021).

Acknowledgments:

We thank Ben Edelman, Mike Luca, Jonathan Roth, Christine Exley, Adam Roberts, Heather Sarsons, The National Fair Housing Alliance, Eva Zschirnt, Brecht Neyt, and Douglas Evans for their helpful feedback.

Funding: None

Author contributions:

Conceptualization: NS, DS
Data curation: NS, DS
Formal Analysis: NS, DS
Investigation: NS, DS
Methodology: NS, DS
Project administration: NS, DS
Software: NS, DS
Supervision: DS
Visualization: NS, DS
Writing – original draft: NS, DS
Writing – review & editing: NS, DS

Competing interests: Dan Svirsky is employed by Uber Technologies, Inc, which did not provide financial support for this project. This project was approved as an "Outside Activity" by Uber Technology's Compliance Organization, but Uber has no oversight of the contents or substance of the paper..

Data and materials availability: Preregistration document, data, and code available here at https://osf.io/zuwa8/

Supplementary Materials

Geographic Variation in Discrimination Measure

We repeat our main experiment, placing each of the 24 ads in the 50 states with an equal budget of $19 across states to assess state-level racial attitudes.

In total, 43,885 Instagram users viewed one of the 1,200 advertisements and sent 5,317 likes. Each state recorded an average of 868 advertisement views and 106 likes for an average like/ad view of 0.12. Results are shown in Table S6 and Figure S1. The 5 states with the lowest $(P_{LZ}-P_{DZ})-(P_L-P_D)$ values -- indicating lower animus towards darker skin complexions -- are Wisconsin, South Carolina, Tennessee, Minnesota, and Utah. The 5 states with the highest $(P_{LZ}-P_{DZ})-(P_L-P_D)$ value are North Carolina, Massachusetts, Florida, Georgia, and Pennsylvania.

We then compare how our metric varies across states to the state-level variance of seven other metrics of race animus. Four are survey based metrics: the Project Implicit self-survey of Black-White racial attitudes, a measurement of racial resentment derived from the American National Election Studies survey (MrP), and two metrics derived from the Cooperative Congressional Election Study (Racial Resentment and Racial Resentment Among Whites). One is a non-survey internet based measure: the popularity of racially charged language on Google search. Another is the number of hate groups in a state. And lastly, we use data on email response rates to Black and White senders from a recent large scale field experiment by Block et al. *(14)*.

Block et al. find that most measures of racial animus are not correlated with each other; each measure may be highlighting a specific independent mechanism of discrimination. Similarly, we find our new metric is not highly correlated with any of these existing measures. These results are shown in Figure S2. All data for these metrics come from the replication data of Block et al.

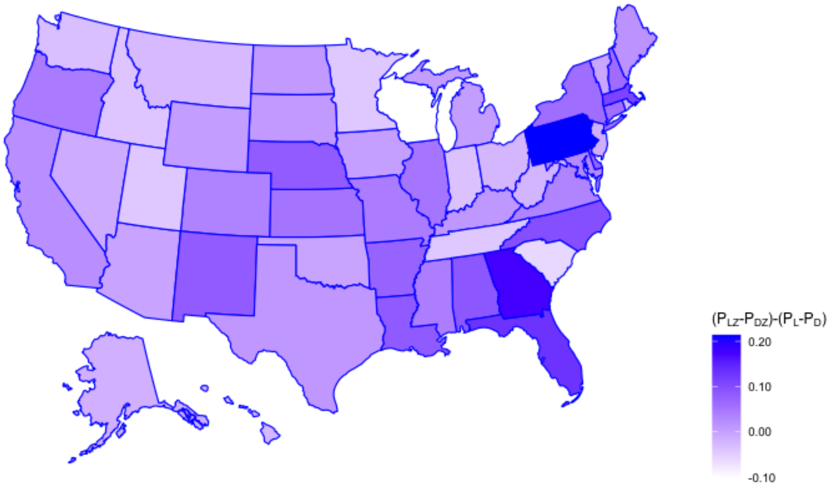


Fig. S1. $(P_{LZ}-P_{DZ})-(P_L-P_D)$ by State. This Figure displays the measure of racial attitudes drawn from our treatment effect by viewer's location at the state level. Darker areas correspond to areas with higher animus towards darker skin complexion advertisements.

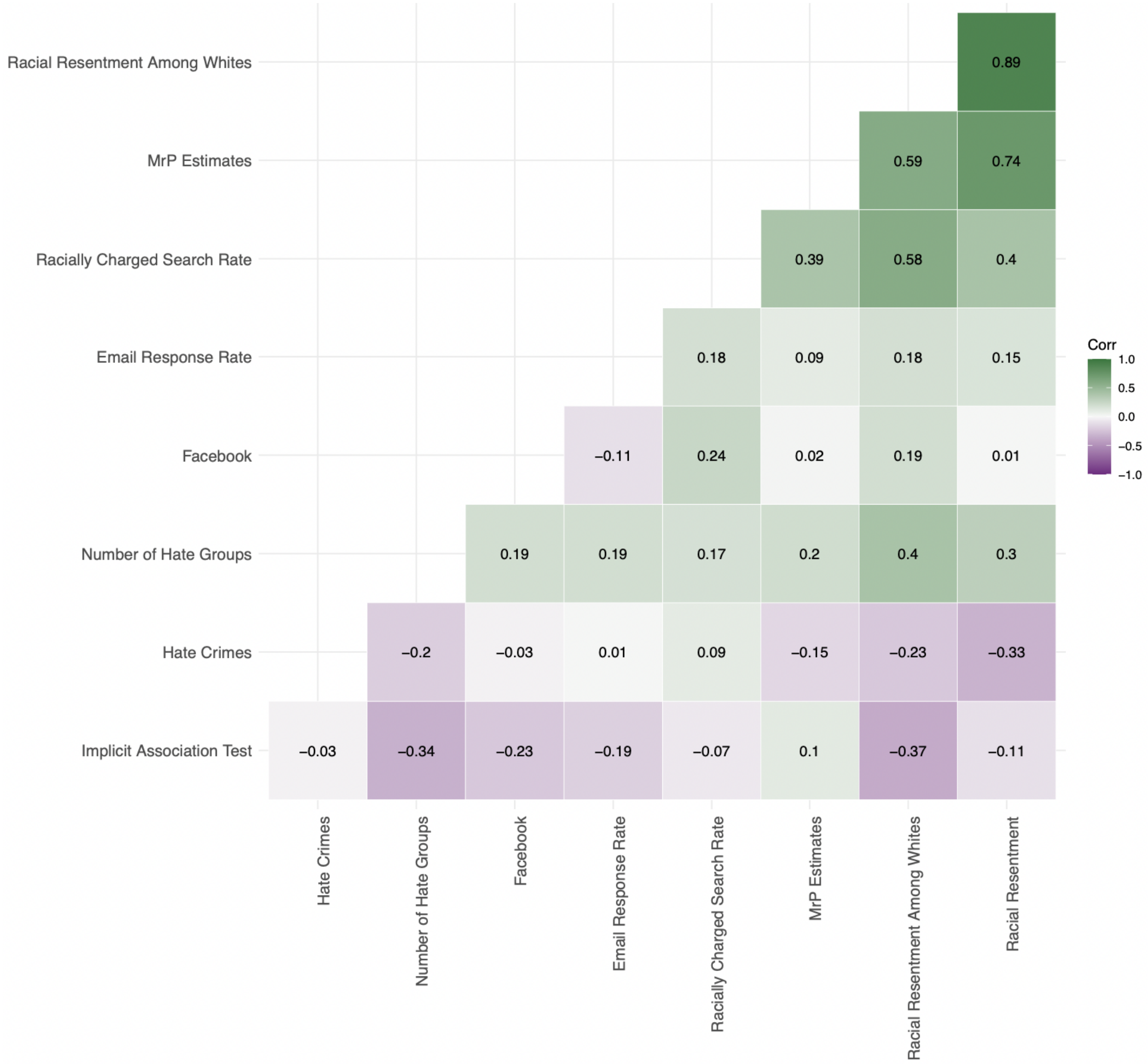


Fig. S2. Measures of Racial Animus by State. This Figure displays various measures of racial attitudes including our treatment effect by viewer's location at the state level. Purple cells correspond to negative correlations and green cells correspond to positive correlations. Block et al. highlight that apart from metrics that are functions of one another such as MrP and Racial Resentment, most measures of racial animus are not strongly correlated.

	Dark Complexion	**Light Complexion**	**Total**
Cropped	1735/7189 (24.1%)	1954/6335 (30.8%)	3689/13524 (27.3%)

Uncropped	1915/10446 (18.3%)	1926/10449 (18.4%)	3841/20895 (18.4%)
Total	3650/17635 (20.7%)	3880/16784 (23.1%)	7530/34419 (21.9%)

Table S1. Likes per Ad Views by Experimental Condition. This figure shows the engagement rate for each experimental group. Specifically, of all advertisement Views, how many times did a viewer “Like” the picture. All comparisons were significant at p<.001 except for Dark Complexion uncropped vs Light Complexion uncropped (p=0.88).

Groups	Image	Likes / Views	Image	Likes / Views	P-value
1	Dark Complexion torso, uncropped	275/1276 (21.6%)	Light Complexion torso, uncropped	240/1479 (16.2%)	<.001
1	Dark Complexion torso, cropped	120/539 (22.3%)	Light Complexion torso, cropped	139/474 (29.3%)	0.01
2	Dark Complexion holding flowers, uncropped	258/1107 (23.3%)	Light Complexion holding flowers, uncropped	271/1162 (23.3%)	0.993
2	Dark Complexion holding flowers, cropped	178/852 (20.9%)	Light Complexion holding flowers, cropped	209/674 (31%)	<.001
3	Dark Complexion holding hands, uncropped	153/483 (31.7%)	Light Complexion holding hands, uncropped	173/709 (24.4%)	0.006
3	Dark Complexion holding hands, cropped	178/938 (19.0%)	Light Complexion holding hands, cropped	159/696 (22.8%)	0.056
4	Dark Complexion keyhole dress, uncropped	450/2665 (16.9%)	Light Complexion keyhole dress, uncropped	490/2711 (18.1%)	0.251
4	Dark Complexion keyhole dress, cropped	312/920 (33.9%)	Light Complexion keyhole dress, cropped	300/711 (42.2%)	<.001
5	Dark Complexion racerback dress, uncropped	416/2445 (17%)	Light Complexion racerback dress, uncropped	427/2772 (15.4%)	0.115

5	Dark Complexion racerback dress, cropped	652/2782 (23.4%)	Light Complexion racerback dress, cropped	834/2887 (28.9%)	<.001
6	Dark Complexion V-back dress, uncropped	363/2470 (14.7%)	Light Complexion V-back dress, uncropped	325/1616 (20.1%)	<.001
6	Dark Complexion V-back dress, cropped	295/1158 (25.5%)	Light Complexion V-back dress, cropped	313/893 (35.1%)	<.001

Table S2. Results Disaggregated by Photo. This Table shows the engagement rate for each experimental advertisement. Specifically, of all advertisement Views, how many times did a viewer “Like” the picture. In the baseline (zoomed out, uncropped) images, we see no statistically significant difference in 3 of the 6 sets, a statistically significant bias towards darker complexions in 2 of the 6 sets, and a statistically significant bias towards the lighter complexion in the remaining 1 set. But in all six cases, when the ads zoom in, the picture with models with darker complexions is penalized more (or improves less), relative to its companion picture with models with lighter complexions.

	Dark Complexion	**Light Complexion**	**Total**
Cropped	1304/11488 (11.4%)	1094/7783 (14.1%)	2398/19271 (12.4%)
Uncropped	1321/11439 (11.5%)	1598/12675 (12.6%)	2919/24114 (12.1%)
Total	2625/22927 (11.4%)	2692/20458 (13.2%)	5317/43385 (12.1%)

Table S3. Aggregated Likes / Ad Views: This table shows the treatment effect for a round of the experiment where the 24 ads were separately run across all 50 states to measure geographic variation. During this experiment, Facebook warned that its audience optimization tools would be turned off because of the number of ads run simultaneously. Compared to the main experimental round, we find lower audience engagement rates, but we still find a treatment effect. All comparisons across complexions were significant at p<.001 except for Dark Complexion uncropped vs Light Complexion uncropped (p=.012).

Ad Name	**Optimization**	**Complexion**	**Cropping**	**Likes/Reach (%)**	**Amount Spent ($)**	**% of dollars spent in group**
flowers	optimized	dark	cropped	9/28 (32.1%)	0.99	2.4
flowers	optimized	dark	uncropped	70/273 (25.6%)	6.16	14.8
flowers	optimized	light	cropped	66/243 (27.2%)	6.98	16.8
flowers	optimized	light	uncropped	359/1368 (26.2%)	27.44	66
flowers	unoptimized	dark	cropped	84/223 (37.7%)	10.37	25
flowers	unoptimized	dark	uncropped	109/458 (23.8%)	10.38	25
flowers	unoptimized	light	cropped	83/218 (38.1%)	10.38	25
flowers	unoptimized	light	uncropped	133/462	10.39	25

				(28.8%)		
hands	optimized	dark	uncropped	11/35 (31.4%)	1.7	4.1
hands	optimized	dark	cropped	152/559 (27.2%)	19.31	46.5
hands	optimized	light	uncropped	68/140 (48.6%)	8.3	20
hands	optimized	light	cropped	96/329 (29.2%)	12.18	29.4
hands	unoptimized	dark	uncropped	63/153 (41.2%)	10.39	25.1
hands	unoptimized	dark	cropped	88/255 (34.5%)	10.31	24.9
hands	unoptimized	light	uncropped	74/150 (49.3%)	10.34	25
hands	unoptimized	light	cropped	84/185 (45.4%)	10.4	25.1
keyhole	optimized	dark	uncropped	184/937 (19.6%)	14.27	34.3
keyhole	optimized	dark	cropped	0/3 (0%)	0.02	0
keyhole	optimized	light	cropped	1/2 (50%)	0.14	0.3
keyhole	optimized	light	uncropped	459/2639 (17.4%)	27.15	65.3
keyhole	unoptimized	dark	cropped	71/157 (45.2%)	10.4	25
keyhole	unoptimized	dark	uncropped	126/552 (22.8%)	10.39	25
keyhole	unoptimized	light	cropped	71/201 (35.3%)	10.35	24.9
keyhole	unoptimized	light	uncropped	128/642 (19.9%)	10.4	25
racerback	optimized	dark	uncropped	1/1 (100%)	0	0
racerback	optimized	dark	cropped	198/921 (21.5%)	9.2	22.2
racerback	optimized	light	cropped	795/2888 (27.5%)	32.21	77.7
racerback	optimized	light	uncropped	1/7	0.05	0.1

				(14.3%)		
racerback	unoptimized	dark	cropped	218/889 (24.5%)	10.38	25
racerback	unoptimized	dark	uncropped	141/606 (23.3%)	10.41	25.1
racerback	unoptimized	light	cropped	238/845 (28.2%)	10.38	25
racerback	unoptimized	light	uncropped	135/649 (20.8%)	10.38	25
torso	optimized	dark	uncropped	65/390 (16.7%)	7.3	17.5
torso	optimized	dark	cropped	0/0 (0%)	0	0
torso	optimized	light	uncropped	374/1784 (21%)	34.3	82.5
torso	optimized	light	cropped	0/0 (0%)	0	0
torso	unoptimized	dark	cropped	49/100 (49%)	10.32	24.9
torso	unoptimized	dark	uncropped	107/558 (19.2%)	10.35	25
torso	unoptimized	light	cropped	63/110 (57.3%)	10.41	25.1
torso	unoptimized	light	uncropped	104/491 (21.2%)	10.34	25
v-back	optimized	dark	uncropped	289/1281 (22.6%)	30.6	73.9
v-back	optimized	dark	cropped	0/18 (0%)	0.15	0.4
v-back	optimized	light	cropped	14/26 (53.8%)	1.16	2.8
v-back	optimized	light	uncropped	96/384 (25%)	9.52	23
v-back	unoptimized	dark	uncropped	98/439 (22.3%)	10.29	24.8
v-back	unoptimized	dark	cropped	73/346 (21.1%)	10.41	25.1
v-back	unoptimized	light	cropped	73/175 (41.7%)	10.35	25
v-back	unoptimized	light	uncropped	87/248 (35.1%)	10.4	25.1

Table S4. Optimization Results Disaggregated by Photo. This Table measures how advertising budget is distributed across different pictures when Facebook's budget optimization feature is turned on or off.

Optimization	**Complexion**	**Cropping**	**Likes**	**Reach**	**Likes/Reach (%)**	**Amount Spent ($)**
Optimized	Dark	Cropped	359	1529	23.48	29.67
Optimized	Dark	Uncropped	620	2917	21.25	60.03
Optimized	Light	Cropped	972	3488	27.87	52.67
Optimized	Light	Uncropped	1357	6322	21.46	106.76
Unoptimized	Dark	Cropped	583	1970	29.59	62.19
Unoptimized	Dark	Uncropped	644	2766	23.28	62.21
Unoptimized	Light	Cropped	612	1734	35.29	62.27
Unoptimized	Light	Uncropped	661	2642	25.02	62.25

Table S5. Optimization Results. This Table measures how advertising budget is distributed across different picture classes when Facebook's budget optimization feature is turned on or off. P-value < 0.001 for Chi-Squared test comparing Amount Spent in first four rows (Optimization OFF) versus Amount Spent in last four rows (Optimization ON).

State	$(P_{LZ}-P_{DZ})-(P_L-P_D)$	Likes	Reach
United States	0.016	5317	43385
Alabama	0.088	111	905
Alaska	-0.019	35	1020
Arizona	-0.005	134	876
Arkansas	0.072	93	869

California	0.02	205	890
Colorado	0.035	120	735
Connecticut	0.057	124	890
Delaware	0.075	57	888
Florida	0.131	189	748
Georgia	0.176	145	783
Hawaii	-0.028	85	1217
Idaho	-0.042	71	970
Illinois	0.05	156	795
Indiana	-0.039	104	856
Iowa	0	78	864
Kansas	0.051	82	869
Kentucky	0.01	100	884
Louisiana	0.083	115	939
Maine	0.015	72	874
Maryland	0.027	141	759
Massachusetts	0.111	124	759
Michigan	0.002	147	729
Minnesota	-0.046	93	804
Mississippi	0.04	106	949
Missouri	0.042	126	881
Montana	-0.029	53	948
Nebraska	0.082	76	995
Nevada	-0.014	98	901
New Hampshire	0.061	72	883
New Jersey	-0.017	174	727
New Mexico	0.082	74	754
New York	0.059	204	783
North Carolina	0.098	155	868

North Dakota	0.007	26	1015
Ohio	-0.026	117	755
Oklahoma	-0.006	93	884
Oregon	0.043	92	707
Pennsylvania	0.214	155	767
Rhode Island	-0.03	71	864
South Carolina	-0.062	119	983
South Dakota	0.005	39	1123
Tennessee	-0.046	147	956
Texas	0.01	179	751
Utah	-0.045	93	741
Vermont	-0.008	38	897
Virginia	0.033	128	796
Washington	-0.036	101	876
West Virginia	-0.023	82	894
Wisconsin	-0.103	86	802
Wyoming	0.002	32	962

Table S6. Discrimination By State. This table shows our measure of discrimination by the viewer's location at the state level, as measured by Facebook. The measure matches our definition of the treatment effect in the experiment: it looks at how much any baseline racial disparities between zoomed-out pictures worsen.