Racial Bias in Customer Service: Evidence from Twitter

Priyanga Gunarathne[a,*], Huaxia Rui[b], Abraham Seidmann[c]

[a] Joseph M. Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA 15260
[b] Simon Business School, University of Rochester, Rochester, NY 14620
[c] Questrom School of Business, Boston University, Boston, MA 02215
*Corresponding author

priyanga.gunarathne@katz.pitt.edu; huaxia.rui@simon.rochester.edu; avis@bu.edu

Abstract

This paper provides the first large-scale evidence of business-to-customer racial bias (B2C bias) on a digital platform, where the perpetrators are individual employees who act on behalf of a company and the victims are customers. This is in contrast to existing studies of racial bias on digital platforms that focus on peer-to-peer marketplaces (e.g., eBay), where both the perpetrators and the victims are individuals acting independently and on their own behalf. In particular, we present the first evidence of B2C bias in corporate social media customer service, a practice that has grown in popularity recently. Unlike traditional call centers, agents providing customer service on social media respond on average to less than half of the complaints they receive, as per our analysis. We investigate the effect of a complaining customer's racial identity, as revealed by the social media profile picture, on the chance of receiving a response.

By analyzing more than 57,000 social media customer complaints to major U.S. airlines and leveraging a variety of analytics techniques, including text mining and facial recognition, we present quantitative evidence that African American customers are less likely to receive a response when they complain, than otherwise similar White customers. Furthermore, our deep-learning-based falsification test shows that the bias is absent without the visual cue that reveals racial identity. This study offers a practical yet powerful recommendation for companies: conceal all customer profile pictures from their employees while delivering social media customer service.

Keywords: social media, customer service, racial bias, deep learning

1 Introduction

Detecting and reporting systemic racial bias is an essential first step toward the ultimate eradication of racial discrimination in our society. Doing so not only requires society members to voice and share their anecdotal experiences, but also relies on researchers to document statistical evidence of racial bias. Previous researchers have documented examples of marketplace discrimination against African American consumers in a multitude of business contexts. A rich terminology, such as "consumer racial profiling" (Gabbidon 2003), "consumer discrimination" (Borjas and Bronars 1989), and "shopping while black" (Schreer et al. 2009), has been used in the literature and by the popular press to describe phenomena that can be broadly defined as less desirable treatment of consumers from a racial minority. This form of racial bias in typically non-digital consumer marketplaces (e.g., restaurants or supermarkets) is essentially *business-to-customer racial bias* (B2C bias), where the perpetrators are individual employees who act on behalf of a company and the victims are customers. The current paper provides the first large-scale evidence of B2C bias on a *digital* platform.

Although several prior studies have investigated racial bias on digital platforms, such as lending platforms (Pope and Sydnor 2011) and ride-sharing platforms (Ge et al. 2016), all these platforms, and the detected racial bias, are peer-to-peer in nature, where perpetrators and victims are mostly individuals operating independently on their own behalf in a sharing-economy marketplace, which is the hallmark of P2P platforms. We call such racial bias peer-to-peer bias (P2P bias). The distinction between P2P bias and B2C bias is not only conceptually important due to the different implications of actions on behalf of individuals vs. actions on behalf of institutions, but also practically important because P2P platforms are unlikely to be held liable for discriminatory behavior by their individual users, such as individual car drivers on Uber or individual sellers on eBay.[1] It is also unlikely that an individual user of a platform will sue

[1] For example, Craigslist maintains a posting service that allows its customers to advertise rental properties, and some customers may post advertisements with clauses such as "*NO MINORITIES*", "*Requirements: Clean Godly Christian Male*", and "*Only Muslims Apply*" that violate the provisions of the Fair Housing Act (FHA). *Chicago Lawyers Committee for Civil Rights Under Law v. Craigslist, 519 F.3d 666 (7th Cir. 2008),* is a Seventh Circuit decision affirming a lower court ruling that Section 230

another individual user for racial discrimination based on statistical evidence derived from platform-wide data or on an isolated incident from a personal account. However, in the case of B2C bias, the business entity is usually vicariously liable, under the *respondeat superior* doctrine, for negligent acts or omissions by its employees in the course of employment.

Clearly, there is an asymmetry in the sense that B2C bias is typically detected in offline contexts but not in digital contexts, while the opposite is true for P2P bias. Such an asymmetry is understandable. Studies of P2P bias in a non-digital context are rare because the high-profile implications of B2C bias naturally attract more attention from researchers. On the other hand, in a digital context, it is much easier to detect P2P bias than to detect B2C bias for two reasons. First, individual users of peer-to-peer platforms are less likely than employees of large organizations to have gone through mandatory discrimination training and hence are more susceptible to implicit or even explicit bias. Second, the fact that there are ample digital footprints of interactions or transactions between companies and their customers should have made companies even more vigilant against potential bias by their employees.

Compared to offline B2C interactions, online B2C encounters have certain unique features that make the study of B2C racial bias on digital platforms a worthwhile endeavor. For instance, while offline B2C racial bias incidents are immediately evident and naturally attract the attention of observers, it is much less likely that someone will detect racial bias in online contexts simply by analyzing individual complaints and their responses, even though consumer-brand interactions on digital platforms may be public (e.g., on social media). Thus, studies such as this one help demystify this task through the use of large-scale data on a digital platform to statistically show the existence of B2C racial bias in an online B2C context. Furthermore, the remedial actions that could mitigate racial bias on digital platforms can be quite different

of the Communications Decency Act (CDA) provides immunity to Craigslist in such a case, because Craigslist is not the speaker of these unlawful statements and does nothing to induce a user to post any particular listing or express any particular preference for discrimination. Similarly, Uber is unlikely to be held legally responsible for discriminatory behavior of its drivers, who are not considered its employees. In 2019, the National Labor Relations Board ruled that Uber drivers are not employees.: https://www.npr.org/2019/05/15/723768986/uber-drivers-are-not-employees-national-relations-board-rules-drivers-saw-it-com and presumably less challenging than in offline B2C contexts. For example, technology itself could be used easily to “blind” the business agents to customer racial profiles on digital platforms, while such a technical solution is infeasible in offline face-to-face B2C encounters.

The current paper presents the first empirical evidence of digitally manifested B2C bias specifically in the context of the private corporate sector (i.e., not government or public service), thereby closing the gap in the literature on racial bias in offline and online contexts. In particular, we demonstrate the existence of racial bias when companies deliver customer service through social media, a practice that has become increasingly popular. Because customers can easily connect with firms through this new channel from anywhere at any time using their smartphones, and because of the asynchronous nature of these interactions, social media customer service has been adopted by almost all major business-to-consumer companies in the United States. Twitter, in particular, has become a popular platform through which firms constantly monitor and respond to their customers. Furthermore, unlike most social media platforms that algorithmically personalize the content feed users see on the platform, Twitter does not block, limit, or remove content, except under specific technical configurations (e.g., protected accounts), technical issues, or undesirable conditions such as abusive or illegal behavior.

Unlike traditional customer service (e.g., call centers), where a customer almost always receives a response as long as he/she is patient, customer complaints directed to firms on social media do not necessarily receive a response. In fact, more than half of the customer complaints in our sample did not receive a response. This phenomenon motivates us to study the following specific research question in order to investigate racial bias in social media customer service:

Research Question: *Are customers of racial minorities less likely to receive a response when they complain to a company’s social media customer service?*

On one hand, there can be at least two underlying mechanisms that might lead to racial discrimination in social media customer service. Some social media customer service agents may simply dislike individuals of certain racial groups and therefore be reluctant to provide service to them on social media (i.e., taste-based discrimination (Becker 1957)). As there is very limited information about the perceived business value of social media users, some customer service agents may also evaluate them based on their racial identities, leading to differential treatment of customers across different racial groups (i.e., statistical discrimination (Arrow 1973; Phelps 1972)). Regardless of the mechanism, the bias may arise explicitly or implicitly, in which case a service agent is not even aware of his/her own behavior. In summary, the availability of digital racial identity information may trigger racial bias in the delivery of customer service on social media, just as physical racial identity information may trigger racial bias in a physical environment.

On the other hand, there are at least two powerful forces that can inhibit the occurrence of racial bias in the context of social media customer service. First, compared to P2P bias, the stakes of B2C bias are significantly higher for any company because racial discrimination has been universally rejected and information dissemination is rapid thanks to social media. While a P2P platform is relatively less vulnerable to public criticism or legal action because of its indirect role in any incident of P2P bias, a business entity involved in a B2C bias incident is directly implicated and responsible for the behavior of its employees and therefore is subject to more public scrutiny and harsher criticism. Moreover, as extensively pointed out in the literature, customer experience plays a major role in shaping the success of any business organization, with important implications for sales, brand loyalty, customer churn/retention rate, and consumer advocacy. Thus, although customer value-based preferential treatment is justified for a profit-oriented business organization, it is inconceivable to have race-based preferential treatment in delivering customer service. Even slight evidence of racial bias in this unique yet crucial component of customer interaction can be detrimental for a brand that wants to thrive in the social media age. Second, unlike the traditional customer service setting, where almost all customer-brand interactions are private, interactions on social media are mostly public. Such an unprecedented level of openness and transparency could make racial bias much less likely to occur. Indeed, sunlight is the best disinfectant, and we believe data transparency, assisted by data analytics, holds the promise to eradicate racial bias.

To empirically address our research question, we obtained a large data set of more than 57,000 customer complaints on Twitter directed explicitly toward the official Twitter accounts of seven major U.S. airlines. We relied on social media profile pictures for racial group identification because previous literature suggests that a visual cue is an important trigger of bias. In particular, behavioral research in psychology suggests that an individual's primitive conscious neural evaluation of another individual's race, which has a consequential impact on the perceiver and the perceived, is usually activated by the stimulus of a human face (Bruce and Young 1986; Calder et al. 2011; Calder and Young 2005). For social media customer service agents, a customer's social media profile picture is the most likely source of such a stimulus. Most previous studies of P2P bias on digital platforms (Pope and Sydnor 2011, Edelman and Luca 2014, Ayres et al. 2015, Younkin and Kuppuswamy 2018) also rely on a visual cue as the trigger of racial bias.

Thanks to the public nature of social media customer service, we are able to control for nearly all of what a social media agent knows about a customer before making the initial response decision. Through matching-based sampling, we find evidence that African American customers (as identified by profile pictures) are less likely to receive brand responses to their complaints on social media than otherwise similar White customers, while Asian and Hispanic customers do not experience such a difference in response rates.

To strengthen the validity of our study, we conduct a falsification test that relies on a proposed deep-learning method for inferring latent demographic attributes of Twitter users from text. The falsification test is based on the premise that a visual cue in the form of human face is usually the trigger of racial bias. The detection of racial bias even in the absence of such a visual cue thus will invalidate our empirical finding. To implement the test, we designed a convolutional neural network to detect a customer's race from the customer's past tweets. We then applied this algorithm to infer the race of those customers whose profile pictures do not contain a human face. Analysis of this sample shows no evidence of racial bias, thereby supporting our main finding and suggesting a strategy to minimize racial bias: adjust social media customer service software so that customer profile pictures are concealed from customer service agents.

The rest of the paper is organized as follows. Section 2 reviews the background literature on racial bias. We present the main analyses in Section 3, the falsification test in Section 4, and some additional analyses in Sections 5 and 6. Section 7 concludes the paper.

2 Literature Background

The psychology literature recognizes bias as a human trait resulting from people's need to classify individuals into categories as they strive to process information fast and understand the external world (Allport 1958). Two conceptualizations of bias, explicit bias and implicit bias, are developed in the literature. With explicit bias, individuals are aware of their prejudices against certain social groups, so their positive or negative preferences are consciously developed (Fridell 2013). In contrast, implicit bias involves subconscious feelings, perceptions, attitudes, and stereotypes that have developed as a result of prior influences and experiences; it can trigger automatic positive or negative preferences toward certain groups and does not require animus (The United States Department of Justice – Police Community Relations Toolkit).

The economics literature distinguishes two models of discrimination. In the first model, known as taste-based discrimination (Becker 1957), developed in the context of the labor market, some employers have a distaste for hiring members of a minority group. This distaste may lead them to refuse hiring members of a minority group, or, if they do hire, to pay them less than other workers for the same level of productivity. The second model, known as statistical discrimination (Arrow 1973; Phelps 1972), views the differential treatment of members of a minority group as the result of a signal extraction problem. In the classic example where an employer assesses the expected productivity of a job candidate (Guryan and Charles 2013), the employer, partially informed of the candidate's productivity, uses the candidate's attributes, such as race, to predict the applicant’s expected productivity, thereby resulting in racial discrimination.

Empirical research on consumer racial discrimination in offline environments focuses exclusively on B2C bias, as is shown in Table 1. For example, Ayres (1991) hired 6 testers (one White female, three White males, one Black female, and one Black male) to negotiate prices at 90 Chicago car dealerships. Prices from 180 negotiations suggested better prices on identical cars for White men than for Blacks. Ondrich et al. (1999) conducted over 1,500 rental housing audits, where each audit consisted of visits to a landlord by a White person and a Black or Hispanic person with similar socio-economic characteristics. They found widespread discrimination across several types of landlord behavior. Schreer et al. (2009) conducted a field experiment where White and Black customers browsed in high-end retail stores and asked a salesperson to remove a security sensor from a pair of sunglasses prior to trying them on in front of a mirror. Although all requests were granted, according to the researchers, the salespersons showed greater levels of suspicion regarding the requests from the Black customers.

As people increasingly shift their social and business activities toward the digital world, researchers have naturally shifted their attention to the detection of racial bias on digital platforms. Pope and Sydnor (2011) examined the online peer-to-peer lending platform Prosper.com and found that loan listings with African Americans in attached pictures were less likely to receive funding than those of Whites with similar credit profiles. Ayres et al. (2015) investigated the impact of seller race in a field experiment involving baseball card auctions on eBay. They found that the cards offered by African American sellers sold for approximately 20% less than cards offered by Caucasian sellers. Ge et al. (2016) examined racial discrimination on ride-sharing platforms such as Uber and Lyft, and they observed longer waiting times and more frequent cancellations for African American passengers. Edelman et al. (2017) investigated the existence of racial discrimination against ethnic minority guests on Airbnb and found that applications from guests with distinctively African American names were 16% percent less likely to be accepted relative to identical guests with distinctively White names. Younkin and Kuppuswamy (2018) examined crowdfunding projects on Kickstarter.com and found that African American men were significantly less likely than similar White founders to receive funding.

Table 1. Racial Bias Studies in Consumer Markets

	B2C	P2P
Physical (Offline)	**Retail:** Schreer et al. (2009) **Housing:** Ondrich et al. (1999), Turner and Mikelsons (1992), Yinger (1986) **Car Sales:** Ayres (1991, 1995), Ayres and Siegelman (1995) **Health Care:** Blair et al. (2013), Penner et al. (2010), Sabin et al. (2008)	
Digital (Online)	Current paper	**Airbnb.com:** Edelman and Luca (2014), Edelman et al. (2017) **Autobytel.com:** Morton et al. (2003) **Craigslist:** Ghoshal and Gaddis (2015), Doleac and Stein (2013) **eBay.com**: Ayres et al. (2015) **Prosper.com:** Pope and Sydnor (2011) **Kickstarter.com:** Younkin and Kuppuswamy (2018) **Uber and Lyft:** Ge et al. (2016)

3 Main Analyses

The ideal design for our research question is to conduct an audit study (i.e., sending the same tweets to different companies and/or with different profile pictures to test the response rates) similar to the field experiment of Edelman et al. (2017). However, such an approach is now considered unethical and unacceptable. A lab experiment cannot fully mirror the conditions of customer service agents responding to customer service complaints on Twitter. As observational studies are appropriate when an experiment is unethical or infeasible (Winship and Morgan 1999), we study our research question using observational data.

We collected all tweets mentioning the official Twitter accounts of seven major U.S. airlines from September 2014 to May 2015, along with the Twitter profile information of the users who sent out these tweets to airlines (e.g., username, location, profile description, profile picture URL, number of followers, number of tweets posted in the past, etc.). To distinguish complaints from all other types of tweets, we developed a lexicon-based complaint classifier with approximately 91% accuracy and F-1 score. The details of this classification process are reported in Section A.1 of Appendix A.

Using this classifier, we obtained 173,662 initial complaining tweets. Then, we removed from this dataset those tweets with technically inaccessible profile picture URLs, which gave us 167,575 complaining tweets for profile image analysis. We passed the profile picture URL associated with each complaining tweet to the Kairos API (www.kairos.com), a commercial cloud API that offers functionalities for face recognition, face identification, face verification, gender/age/race detection, and multi-face detection. The Kairos API could successfully detect a face in profile pictures associated with 110,533 complaining tweets in our sample. Upon detecting a face, the Kairos API reports the probabilities that the individual in the image is White, African American, Hispanic, Asian, or from another racial category. To ensure that all tweets in our analysis are from users whose profile pictures indicate unambiguously a certain race, we assigned a user to a particular racial category only if the user was assessed by Kairos API a probability of 0.9 or greater for that racial category. Moreover, the gender of each detected face was also obtained from the Kairos API. This process resulted in a sample of 59,984 complaining tweets for further analysis[2].

About 4% of this sample were complaining tweets from users with “verified” status on Twitter. At the time we collected the data, Twitter granted the “verified” status to celebrity users in music, acting, fashion, government, politics, religion, journalism, media, sports, business, and other key interest areas. As any insight that is driven by this special group of users may not generalize to ordinary customers, we dropped this small group of users from our sample so that the results are completely driven by ordinary

[2] A Chi Squared test of homogeneity shows that there is no systematic complaint misclassification across racial groups. See Section A.3 of Appendix A for details.

users. As a result, our final sample comprises 57,484 complaining tweets. Supplementary statistics and details on the data preparation process are presented in Sections A.1-6 in Appendix A. Table 2 reports the number of complaints and the response rates for each racial group in our sample.

Table 2. Response Rates by Race

Racial Category	No. of Complaints	Percentage (%)	Response Rate (%)
White	48,843	84.97	49.50
African American	4,511	7.85	44.78
Asian	3,645	6.34	50.07
Hispanic	430	0.75	50.70
Other	55	0.10	40.00
Total	**57,484**		

We list in Table 3 the definitions of all key variables used in the empirical analysis and report their summary statistics in Table 4.

To conduct empirical analyses, we created three matched samples, each corresponding to one minority racial group (i.e., African American, Asian, and Hispanic). For example, for the African American sample, the treatment group consists of complaining tweets from African American users, and the control group consists of matched complaining tweets selected from White users. We use Propensity Score Matching (PSM) to match each complaining tweet in the treatment group with at most one complaining tweet from the control group with no replacement and common support, using all the observable covariates. For all the matched samples, the absolute standardized percentage bias of all covariates is below 10%, suggesting adequate balance. Covariate imbalance check summary statistics for the matched samples are reported in Table 5. Graphical summaries of the covariate imbalance analyses and detailed statistics are reported in Section A.7 of Appendix A.

Table 3. Definitions of Variables

Variable	Definition
Responded	1 if the airline responded to the complaining tweet, 0 otherwise
User Characteristics	
African American	1 if the customer is African American, 0 otherwise
Asian	1 if the customer is Asian, 0 otherwise
Hispanic	1 if the customer is Hispanic, 0 otherwise
White	1 if the customer is White, 0 otherwise
Other	1 if the customer is not African American/Asian/Hispanic/White, 0 otherwise
Female	1 if the customer is female, 0 otherwise
Followers	Number of followers the user had, at the creation of the complaining tweet. Log transformed as *log(followers+1)*
Updates	Number of tweets ever posted by the user (log transformed)
Public Profile	1 if the user's location, website, or profile description (i.e., Twitter bio) is publicly available, 0 otherwise
Text Complexity	
Length in Words	Number of words in the tweet
Syllables per Word	Number of syllables per word for the tweet
Characters per Word	Number of characters per word for the tweet
Positive Intensity	Positive sentiment intensity of the tweet per word, based on the AFINN sentiment lexicon
Negative Intensity	Negative sentiment intensity of the tweet per word, based on the AFINN sentiment lexicon
Offensive	1 if the complaining tweet contains offensive words, 0 otherwise
Slang	1 if the complaining tweet contains slang, 0 otherwise
Multiple Users Mentioned	1 if multiple user accounts are mentioned in the complaining tweet, 0 otherwise
Hashtag	1 if the complaining tweet contains hashtags, 0 otherwise
URL	1 if the complaining tweet contains web URLs, 0 otherwise
Smileys	1 if the complaining tweet contains smileys, 0 otherwise
Order	The position of the airline Twitter handle in the complaining tweet, relative to other username mentions, if any
Twitter Handle First	1 if the complaining tweet starts with a Twitter user handle, 0 otherwise
Cluster	Categorical variable indicating the cluster ID assigned to the complaining tweet based on text clustering performed on the tweets
Other Tweet-based Variables	
Complaints within the Previous Hour	Number of complaining tweets received by the airline during the hour prior to receiving the current complaining tweet (log transformed)
Retweets	Number of times the tweet was retweeted before the first response from the airline (if the airline responded), or before the end of the observation period (if the airline did not respond), log transformed
Other	
Day of Week	Categorical variable indicating the day of the week
Airline	Categorical variable indicating the airline to which the user tweet was sent

Table 4. Summary Statistics			
Variable	**Observations**	**Mean**	**Std. Dev.**
Responded	57,484	0.4917	0.4999
African American	57,484	0.0785	0.2689
White	57,484	0.8497	0.3574
Asian	57,484	0.0634	0.2437
Hispanic	57,484	0.0075	0.0862
Other	57,484	0.0010	0.0309
Female	57,484	0.4649	0.4988
Followers	57,484	5.4303	1.7537
Updates	57,484	7.1424	2.1885
Public Profile	57,484	0.8966	0.3044
Length in Words	57,484	19.1088	5.6948
Syllables per Word	57,484	1.3710	0.4918
Characters per Word	57,484	4.5159	0.8713
Positive Intensity	57,484	0.0478	0.0906
Negative Intensity	57,484	0.1879	0.2424
Offensive	57,484	0.0268	0.1615
Slang	57,484	0.0252	0.1569
Multiple Users Mentioned	57,484	0.2826	0.4503
Hashtag	57,484	0.2464	0.4309
URL	57,484	0.0780	0.2682
Smileys	57,484	0.0178	0.1323
Order	57,484	2.0639	0.8417
Twitter Handle First	57,484	0.6185	0.4858
Retweets	57,484	0.0410	0.2081
Complaints within the Previous Hour	57,484	2.3921	0.9839

Table 5. Covariate Imbalance Check Summary Statistics for the Matched Samples									
Variable	African American			Asian			Hispanic		
	Mean		Std. % Bias*	Mean		Std. % Bias*	Mean		Std. % Bias *
	Treated	Control		Treated	Control		Treated	Control	
Female	.53385	.54695	-2.6	.54089	.53952	0.3	.42326	.44186	-3.7
Followers	5.677	5.5512	7.4	5.2126	5.2015	0.6	5.353	5.2873	4
Updates	8.1669	8.0185	6.7	7.0542	7.026	1.3	7.3872	7.3179	3.3
Public Profile	.9232	.91632	2.4	.8949	.89544	-0.2	.89302	.91628	-7.5
Length in Words	18.399	18.334	1.1	18.954	18.802	2.6	18.284	18.047	4
Syllables per Word	1.3676	1.3714	-0.8	1.3609	1.3686	-1.6	1.3628	1.3907	-5.7
Characters per Word	4.5261	4.5303	-0.5	4.5176	4.551	-2.9	4.4953	4.4977	-0.3
Positive Sentiment Intensity per Word	.04578	.044	1.9	.0466	.04607	0.6	.04963	.04566	4.2
Negative Sentiment Intensity per Word	.21305	.22699	-5.5	.19375	.19921	-2.1	.22395	.22886	-1.9
Offensive	.04195	.05216	-5.7	.03375	.03458	-0.5	.03953	.03721	1.3
Slang	.05172	.06016	-4.5	.03622	.04199	-3.4	.04186	.04884	-4
Multiple Users Mentioned	.28746	.28058	1.5	.24835	.26043	-2.7	.28372	.26512	4.1
Hashtags	.19179	.19933	-1.8	.23024	.236	-1.3	.23953	.22558	3.2
URL	.06326	.06326	0	.08919	.0848	1.6	.07209	.06047	4.4
Smileys	.01199	.0111	0.7	.0225	.02305	-0.4	.01628	.0093	5.4
Order	2.1356	2.1352	0.1	1.983	2.0027	-2.3	2.1256	2.0744	6.2
Twitter Handle First	.58091	.58779	-1.4	.59687	.59138	1.1	.62093	.62326	-0.5
Retweets	.0362	.03701	-0.4	.0341	.04042	-3.1	.02497	.02269	1.2
Complaints within the Previous Hour	2.3949	2.4101	-1.6	2.3373	2.3344	0.3	2.4482	2.4514	-0.3
American Airlines (vs. Alaska)	.31521	.32719	-2.6	.2489	.24067	1.9	.37674	.3907	-3
Delta (vs. Alaska)	.11632	.11188	1.4	.09715	.09468	0.8	.07442	.06047	4.8
JetBlue (vs. Alaska)	.11521	.11498	0.1	.09742	.10675	-3.2	.08837	.07907	3.2
SouthWest (vs. Alaska)	.18224	.18335	-0.3	.13666	.13611	0.2	.16279	.17209	-2.6
United (vs. Alaska)	.21931	.2111	1.9	.31229	.3112	0.2	.22558	.23721	-2.7
Virgin America (vs. Alaska)	.03529	.03263	1.4	.06778	.07217	-1.9	.04884	.03721	5.6
1.cluster	.01398	.01287	0.9	.01729	.01647	0.6	.01163	.0093	1.9
2.cluster	.02064	.02109	-0.3	.01427	.01537	-0.9	.03488	.02791	4.4
3.cluster	.01398	.01154	2	.01701	.01454	1.9	.02326	.01163	8.2
4.cluster	.01731	.01754	-0.2	.01894	.01619	2	.02326	.02558	-1.6
5.cluster	.0222	.02375	-1	.02223	.02333	-0.7	.03256	.04884	-9.7
6.cluster	.05372	.04861	2.2	.08233	.07903	1.3	.05814	.03953	7.8
7.cluster	.04084	.0404	0.2	.04363	.04171	0.9	.02791	.0186	5

8.cluster	.01687	.01731	-0.4	.01921	.01894	0.2	.03256	.03256	0
9.cluster	.02309	.03041	-5.6	.01756	.01701	0.5	.02093	.01628	3.7
10.cluster	.03707	.03263	2.4	.05049	.0579	-3.6	.03023	.03023	0
11.cluster	.02242	.02131	0.8	.02195	.01839	2.4	.0186	.01395	3.3
12.cluster	.01842	.01798	0.3	.02113	.0225	-0.9	.02326	.03023	-4.5
13.cluster	.01798	.02153	-2.6	.02333	.02278	0.4	.0186	.01628	1.7
14.cluster	.04107	.04639	-2.9	.03238	.03128	0.6	.05349	.04884	2.3
15.cluster	.01176	.01354	-1.7	.00878	.01125	-2.5	.01163	.00698	4.5
16.cluster	.09723	.08901	2.8	.08754	.09413	-2.3	.06279	.07209	-3.5
17.cluster	.02952	.0293	0.1	.02442	.02634	-1.2	.02093	.03488	-9.1
18.cluster	.01709	.01865	-1.3	.01537	.01647	-0.9	.01163	.0093	2.1
19.cluster	.01976	.01776	1.4	.02195	.0225	-0.4	.0093	.0093	0
20.cluster	.04218	.04306	-0.5	.02744	.02799	-0.3	.03023	.03953	-5.5
21.cluster	.03219	.03152	0.4	.02854	.02525	2	.02791	.03023	-1.4
22.cluster	.0162	.01998	-3.1	.01564	.01482	0.7	.01395	.0186	-4
23.cluster	.01398	.0131	0.7	.02278	.01701	3.9	.0186	.01628	1.7
24.cluster	.01643	.01754	-0.9	.01454	.01153	2.6	.00465	.00698	-2.5
25.cluster	.01709	.01421	2.1	.01921	.02415	-3.6	.01628	.01628	0
26.cluster	.01931	.01665	1.9	.01839	.02031	-1.4	.02791	.03023	-1.5
27.cluster	.01509	.01532	-0.2	.0118	.014	-2.2	.01628	.0186	-2.1
28.cluster	.01421	.01354	0.5	.01537	.01619	-0.6	.01163	.0186	-5.7
29.cluster	.03529	.03085	2.4	.02744	.02799	-0.3	.02791	.02558	1.3
30.cluster	.01287	.01354	-0.5	.01674	.01811	-1	.02558	.01163	9.2
31.cluster	.05971	.0677	-3.7	.0365	.03732	-0.4	.03488	.04884	-7.5
32.cluster	.03751	.03818	-0.3	.05543	.05214	1.5	.03953	.03721	1.2
33.cluster	.0182	.01909	-0.7	.02031	.01866	1.2	.03256	.02093	7.3
34.cluster	.01909	.01731	1.2	.01509	.01647	-1	.03721	.04419	-4
35.cluster	.01043	.0091	1.3	.00549	.00604	-0.6	.00698	.00465	2.4
36.cluster	.01643	.01487	1.2	.01427	.01482	-0.4	.01163	.01163	0
37.cluster	.01598	.01776	-1.2	.02552	.02717	-1	.02326	.03256	-6
38.cluster	.01731	.01842	-0.9	.01427	.01262	1.4	.02326	.0186	3.5
39.cluster	.02508	.02397	0.8	.01619	.01317	2.3	.02791	.03256	-3.1
2.day of week	.16293	.15361	2.5	.17124	.16877	0.7	.17442	.20698	-8.5
3.day of week	.14606	.15006	-1.1	.14133	.13721	1.2	.14186	.12791	4
4.day of week	.13518	.12986	1.6	.132	.132	0	.12093	.10698	4.2
5.day of week	.14273	.14495	-0.6	.13858	.14023	-0.5	.13256	.1186	4.1
6.day of week	.1465	.14606	0.1	.14682	.14874	-0.5	.15581	.16047	-1.3
7.day of week	.1374	.14184	-1.3	.13611	.13804	-0.6	.1093	.09535	4.3

Std. % Bias*: The standardized percentage bias, which is the percentage difference of the sample means in the treated and non-treated sub-samples as a percentage of the square root of the average of the sample variances in the treated and non-treated groups.

The econometric specification is a logistic regression where the dependent variable equals one if the complaint receives a response from the airline and zero otherwise, based on Twitter metadata. The independent variable is the binary race indicator. We include a variety of control variables, such as the text cluster fixed effects to account for the topic and style of each complaint (Section A.4 of Appendix A), linguistic attributes related to the complexity of the tweet, tweet traffic, retweets, and various user characteristics. Table 6 presents the estimation results from our main empirical specification.

Column (1) of Table 6 reports results from a pooled regression on the initial dataset. The coefficient for *African American* is negative and statistically significant, suggesting a lower response rate for African American customers than for White customers. Columns (2)-(7) of Table 6 report the coefficient estimates for both the matched dataset and the full dataset comprising each minority racial group and White customers (i.e., PSM vs. All). The PSM-based estimate in column (2) for *African American* is negative and statistically significant, again suggesting that African American customers are less likely to receive responses to their complaints than White customers who are otherwise similar. Specifically, being an African American customer decreases the odds of receiving a response from the airlines by approximately 12% compared with similar White customers, holding all other variables constant. Therefore, our analysis reveals racial bias against African American customers, while such a bias is not evident for Asian or Hispanic customers.

To examine the possibility that the observed racial bias against African Americans is driven by some unobservable confounding factors such as potential systematic differences between African Americans and non-African Americans (e.g., flying very different routes and paying different airfares), we performed some additional heterogeneity checks, which are presented in Section A.9 of Appendix A. We did not find any statistically significant difference between African Americans and non-African Americans in terms of their travel patterns that could potentially bias our results.

Table 6. Estimation Results

Variables	Pooled (1)	African American and White		Asian and White		Hispanic and White	
		PSM (2)	All (3)	PSM (4)	All (5)	PSM (6)	All (7)
African American *(baseline: White)*	**-0.1231*****	**-0.1181****	**-0.1258*****				
	(0.0357)	**(0.0471)**	**(0.0358)**				
Asian *(baseline: White)*	0.0123			0.0052	0.0135		
	(0.0381)			(0.0517)	(0.0382)		
Hispanic *(baseline: White)*	0.0798					0.1899	0.0860
	(0.1040)					(0.1687)	(0.1045)
Other *(baseline: White)*	-0.5678*						
	(0.2952)						
Female	-0.0107	-0.0534	-0.0121	0.0241	-0.0024	0.2141	-0.0019
	(0.0187)	(0.0479)	(0.0194)	(0.0524)	(0.0196)	(0.1714)	(0.0203)
Observations	57,484	9,010	53,354	7,288	52,488	860	49,273
Log Likelihood	-34177.17	-5300.866	-31696.73	-4366.387	-31199.21	-435.4723	-29220.76
AIC	68504.34	10745.73	63537.47	8876.774	62542.42	1014.945	58585.53
BIC	69176.28	11257.37	64177.17	9373.141	63180.94	1357.444	59219.5

NOTE: For brevity, estimated results for a set of selected variables are reported. For all results, see Section A.8 of Appendix A.
*** p<0.01, ** p<0.05, * p<0.1 (Robust standard errors in parentheses)

4 Falsification Test

In our main analysis, we essentially compared the response rates for customers belonging to a minority racial group and for similar White customers, where racial identity is revealed by profile images. The underlying premise is that the racial bias from social media customer service agents, if any, is most likely triggered by a race-revealing visual cue. Our falsification test relies on the counterfactual of this premise: if our econometric analysis can nevertheless detect the presence of racial bias even for a sample of customers whose racial identity is not directly visible to social media customer service agents, then our previously detected racial bias is likely driven by some unobservable confounding factors, thereby invalidating our empirical findings.

To implement this falsification test, we focus on complaining customers whose profile pictures are classified as "No faces detected" (e.g., pictures of pets, symbols, Twitter's default profile picture for people with no profile picture, etc.) by the Kairos API. For these customers, there is no easy way, if any at all, for a social media customer service agent to infer their racial identities using profile images. After removing tweets from verified users, we obtain 43,048 complaining tweets eligible for this falsification test.

Of course, the key challenge of this falsification test is for us researchers to know the racial identities of customers in the absence of their facial information. To meet this challenge, we developed a deep-learning classification method to infer the race of a Twitter user (i.e., African American or not) based on the user's historical tweets. The proposed method is detailed in Appendix B. We then inferred the racial identities of the users in our falsification test sample and identified 1,107 complaining tweets from African American users and 41,941 complaining tweets from non-African American users.

We first estimated our benchmark specification on this falsification data sample. The coefficient estimates are reported in column (2) of Table 7. We find that the coefficient of *African American* is not statistically significant. In other words, the racial bias we previously detected does not exist as long as a social media agent cannot infer a customer's race from the customer's profile picture. This result suggests that the detected racial bias against African Americans is unlikely to be a spurious correlation.

Next, similar to the approach in our main analysis, we constructed a subset of the falsification test sample where the treatment group consists of complaining tweets from African American users and the control group consists of matched complaining tweets selected from non-African American users, using all the observable covariates and with no replacement and common support in matching. Covariate imbalance check analyses for this matching procedure are reported in Section A.10 of Appendix A. Column (1) of Table 7 reports the falsification test results using this subsample; these results are consistent, thereby further supporting our main finding.

Table 7. Estimation Results — Falsification Test

Variable	**(1)** **PSM**	**(2)** **All**
African American (*baseline: non-African American*)	0.0636 (0.1022)	0.0340 (0.0703)
Female	-0.0620 (0.1384)	0.0336 (0.0299)
Observations	2,212	43,048
Log Likelihood	-1161.827	-24756.58
AIC	2467.655	49657.16
BIC	2878.174	50281.4

For brevity, estimated results for a set of selected variables are reported. See Section A.11 in Appendix A for details.
*** p<0.01, ** p<0.05, * p<0.1 (Robust standard errors in parentheses)

To examine whether the results from our main analysis are comparable with the results from the falsification test, we analyzed the tweets from users with profile pictures and from users without profile pictures. Specifically, we generated word clouds for tweets written by the two groups of users (i.e., those with profile photos and those without) and also their text-attribute-based summary statistics. We find these word clouds and summary statistics as reported in Section A.12 of Appendix A, are largely comparable albeit not identical. We acknowledge this as a limitation of our proposed falsification testing approach.

5 Potential Mitigators of Racial Bias

In this section, we examine two factors that might mitigate racial bias against African Americans in the social media customer service context. As described earlier, statistical discrimination occurs when there is insufficient information available for the decision maker to assess the true value of the individuals involved, so a decision is (unconsciously) made based on protected attributes such as race. The underlining conjecture is that any factor that enhances the perceived value of a customer in the eyes of the decision maker could effectively reduce any racially biased treatment of that customer. To empirically investigate this, we first examined whether the social media popularity of the complaining customer, as measured by the number of followers, could effectively enhance the perceived value of African American customers such that the discriminatory treatment that we witnessed for African Americans is less evident for Twitter users with greater social media popularity. We re-estimated the African American-White sample from our main analysis with the additional interaction term *African American*Followers* and find that this interaction is not statistically significant in the PSM-matched sample. Therefore, the analysis does not provide empirical evidence to support statistical discrimination as an underlying mechanism of racial bias in the social media customer service context, at least when examined in the lens of social media popularity.

Another potential mitigating factor could be the audience size of a complaining tweet, as the exposure of a complaint to a larger audience might overpower any racially biased, potentially unconscious judgements. One factor that affects the audience size of a tweet is whether the tweet begins with a user mention (e.g., @airline my flight has been delayed for 2 hours!). In other words, although a tweet is generally viewable by all the followers of the author of that tweet, having a Twitter handle at the beginning of a tweet restricts the audience size of that tweet to only a few parties, such as the author, the recipient, and the followers of both the author and the recipient. Again, we re-estimated the African American-White sample from our main analysis, with the additional interaction term *African American * Twitter Handle First*. We find that this interaction term is not statistically significant in the matched sample. Therefore, the analysis does not provide empirical evidence that audience size can mitigate racial bias against African American customers in the social media customer service context. Summary results of these two analyses are reported in Section A.13 of Appendix A.

6 Alternative Outcome Variables

In our main analysis, we have focused on whether a customer's complaint received a response, as it is the most likely metric through which we can detect discriminatory treatment. In this section, we check two other outcome measures (*time to response* and the *overall sentiment of the reply*) that could potentially offer alternative avenues to examine racial bias in the social media customer service context.

To examine the role of a customer’s race on the time required to receive a response (i.e., *Time-to-Response*), we estimated a series of survival analysis models using semi-parametric survival modeling techniques. For a tweet that received a reply, the survival time is defined as the interval between the creation of the tweet and the receipt of the response. The analysis followed the same design as the benchmark specification (Table 6), which includes a pooled regression and a matched-sample analysis for each minority racial group. We estimated a Cox proportional hazards model for each dataset. The detailed estimation results are reported in Section A14 of Appendix A. We do not find statistically significant difference between White Americans and racial minority groups in terms of the time-to-response outcome variable. One plausible explanation for this finding is that implicit racial bias is subconscious and instant, so its effect on customer service agents lasts just long enough to affect the initial binary decision whether to respond, but not long enough to impact the response time.

Next, we examine the overall sentiment of the response as an additional outcome variable. To compute the overall sentiment of airline replies, we use VADER (Valence Aware Dictionary and Sentiment Reasoner), a lexicon- and rule-based sentiment analysis software that is specifically attuned to sentiments expressed on social media. To measure the overall sentiment of a reply tweet sent by an airline, we use VADER's "compound score," which is the most useful metric when a single unidimensional measure of sentiment for a given text is needed.

We estimated an OLS regression model following the same design as the benchmark specification. The detailed estimation results are reported in Section A.15 of Appendix A. We do not find statistically significant evidence of racial bias in terms of the overall sentiment of the replies. This is not surprising, because the standard training of customer service agents on how to craft their replies to customers means there is limited variation in reply sentiment.

7 Conclusion

Using a unique dataset of customer complaints on Twitter to seven major U.S. airlines over a period of nine months, and leveraging facial recognition and deep learning techniques, we investigate the effect of a customer's racial identity, as signaled by the profile picture, on the chance of receiving a response when he or she complains to an airline's social media customer service. The evidence is clear that airlines are less likely to respond to complaints from African American customers than to those from similar White customers, while customers of other racial minorities do not experience such a difference.

The contribution of the current paper to the literature is twofold. First, to the best of our knowledge, this is the first study to empirically investigate the existence of B2C racial bias on a digital platform, which is substantially different from the other reported incidents of racial bias in the P2P context. Therefore, this study closes the gap between the literature on racial bias in digital contexts and the traditional literature on racial bias in offline contexts where the bias is B2C.

Second, our falsification test, which effectively leverages deep learning techniques and social media data to predict latent attributes (i.e., race and gender) of social media users, can be useful for social science researchers when the key independent variable is derived from images and may be endogenous.

Our findings have important implications for practitioners, especially for companies that are trying to harness the power of social media to deliver customer service. The empirical results provide evidence suggesting racial bias against African American customers in social media customer service. Given the particularly severe consequences of B2C racial bias, it is of vital importance that companies carefully examine the root causes of such bias so that they can take appropriate actions accordingly and promptly. Whether the bias is implicit or explicit, the company has the responsibility to be aware of the discriminatory aspects of behavior resulting from these biases (Holroyd 2015).

The challenges companies face in the battle against racial discrimination can be quite distinct from those in traditional customer service settings. For example, due to automatic call routing mechanisms, customer service agents in traditional call centers do not have much discretion over which customers they respond to, while offline in-person B2C encounters are mostly private. However, in a social media customer service setting where a large number of interactions takes place online every day and where systematic monitoring and quality assurance practices may not be in place, data analytics-driven audit programs and algorithmically fair automatic routing mechanisms can be helpful to tackle explicit biases, in addition to traditional discrimination prevention programs for social media customer service teams. As implicit human biases are difficult to eradicate, our study strongly recommends that companies adjust their social media customer service software so that customer profile pictures are hidden from their customer service agents, which should go a long way toward preventing potential implicit bias against racial minorities.

References

Allport G W (1958) *The Nature of Prejudice* (Abridged Garden City, N.Y., Doubleday).

Arrow K (1973) The theory of discrimination. *Discrimination in Labor Markets* 3(10): 3-33.

Ayres I (1991) Fair driving: Gender and race discrimination in retail car negotiations. *Harvard Law Review* (104): 817-72.

Ayres I (1995). Further evidence of discrimination in new car negotiations and estimates of its causes. *Michigan Law Review* 94(1): 109-47.

Ayres I, Banaji M, Jolls, C (2015) Race effects on eBay, *The RAND Journal of Economics* 46(4): 891-917.

Ayres I, Siegelman P (1995) "Race and gender discrimination in bargaining for a new car. *American Economic Review* (85): 304-321.

Becker G S (1957) *The Economics of Discrimination* (University of Chicago press).

Blair I V, Havranek E P, Price D W, Hanratty R, Fairclough D L, Farley T, Hirsh H K, Steiner J F (2013) Assessment of biases against Latinos and African Americans among primary care providers and community members. *American Journal of Public Health* 103(1): 92-98.

Borjas G J, Bronars S G (1989) Consumer discrimination and self-employment. *Journal of Political Economy* 97(3): 581-605.

Bruce V, Young A (1986) Understanding face recognition. *British Journal of Psychology* 77(3): 305-327.

Calder A, Rhodes G, Johnson M, Haxby J eds. (2011) *Oxford handbook of face perception* (Oxford University Press)

Calder A J, Young A W (2005) Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience* 6(8): 641-651.

Doleac J L, Stein L C (2013) The visible hand: Race and online market outcomes. The *Economic Journal* 123(572): F469-F492.

Edelman B G, Luca M (2014) Digital discrimination: The case of Airbnb.com. *Harvard Business School NOM Unit Working Paper*: 14-054.

Edelman B, Luca M, Svirsky D (2017) Racial discrimination in the sharing economy: evidence from a field experiment. *American Economic Journal: Applied Economics* 9(2): 1-22.

Fridell L (2013) This is not your grandparents' prejudice: the implications of the modern science of bias for police training. *Translational Criminology* 5: 10-11.

Gabbidon S L (2003) Racial profiling by store clerks and security personnel in retail establishments: An exploration of "Shopping While Black. *Journal of Contemporary Criminal Justice* 19(3): 345-364.

Ge Y, Knittel C R, MacKenzie D, Zoepf S (2016) Racial and gender discrimination in transportation network companies. Working Paper (No. w22776), National Bureau of Economic Research.

Ghoshal R, Gaddis S M (2015) Finding a roommate on Craigslist: Racial discrimination and residential segregation. Available at SSRN 2605853.

Guryan J, Charles K K (2013) Taste-based or statistical discrimination: The economics of discrimination returns to its roots. *The Economic Journal* 123(572).

Holroyd J (2015) Implicit bias, awareness, and imperfect cognitions. *Consciousness and Cognition* (33): 511–523.

Morton F S, Zettelmeyer F, Silva-Risso J (2003) Consumer information and discrimination: Does the internet affect the pricing of new cars to women and minorities? *Quantitative Marketing and Economics* 1(1): 65-92.

Ondrich J, Stricker A, Yinger J (1999) Do landlords discriminate? The incidence and causes of racial discrimination in rental housing markets. *Journal of Housing Economics* (8): 185-204.

Penner L A, Dovidio J F, West T V, Gaertner S L, Albrecht T L, Dailey R K, and Markova T (2010) Aversive racism and medical interactions with black patients: A field study. *Journal of Experimental Social Psychology* 46(2): 436-440.

Phelps E S (1972) The statistical theory of discrimination. *American Economic Review* 62(4): pp.659-661.

Pope D G, Sydnor J R (2011) What's in a picture? Evidence of discrimination from prosper. Com *Journal of Human resources* 46(1): 53-92.

Sabin J A, Rivara F P, Greenwald A G (2008) Physician implicit attitudes and stereotypes about race and quality of medical care. *Medical Care:* 678-685.

Schreer G E, Smith S, Thomas K (2009) Shopping while black: Examining racial discrimination in a retail setting. *Journal of Applied Social Psychology* 39(6): 1432-1444.

The United States Department of Justice, Understanding Bias: A Resource Guide. *Police-Community Relations Toolkit*. Available at https://www.justice.gov/crs/file/836431/download

Turner M A, Mikelsons M (1992) Patterns of racial steering in four metropolitan areas. *Journal of Housing Economics* 2(3):199-234.

Winship, C, and Morgan, S L (1999). The estimation of causal effects from observational data. Annual Review of Sociology 25(1): 659–706.

Yinger J (1986) Measuring discrimination with fair housing audits: Caught in the act. *American Economic Review* 76(5): 881-893.

Younkin P, Kuppuswamy V (2018) The colorblind crowd? Founder race and performance in crowdfunding," *Management Science* 64(7): 3269-3287.