

The Unintended Consequences of Raising Awareness: Knowing About the Existence of Algorithmic Racial Bias Widens Racial Inequality

Shunyuan Zhang
Yang Yang

Working Paper 22-017



The Unintended Consequences of Raising Awareness: Knowing About the Existence of Algorithmic Racial Bias Widens Racial Inequality

Shunyuan Zhang
Harvard Business School

Yang Yang
University of Florida, Warrington College of Business

Working Paper 22-017

Copyright © 2021 by Shunyuan Zhang and Yang Yang.

Working papers are in draft form. This working paper is distributed for purposes of comment and discussion only. It may not be reproduced without permission of the copyright holder. Copies of working papers are available from the author.

Funding for this research was provided in part by Harvard Business School.

The Unintended Consequences of Raising Awareness: Knowing About the Existence of Algorithmic Racial Bias Widens Racial Inequality

Abstract

In May 2016, a series of news articles on algorithmic racial bias went viral and triggered a nationwide outrage on social media, raising public awareness of the existence of algorithmic racial bias. The new public awareness raised an important question: How does awareness of the existence of racial bias in some algorithms influence the disadvantaged group's receptivity to unrelated algorithms that are unbiased (i.e., race-blind) and offer significant financial, education, or health benefits? The current research is the first attempt to investigate this question. An analysis of an Airbnb dataset reveals that, before the 2016 media coverage of algorithmic racial bias, Black hosts were less likely than white hosts to use Smart Pricing (i.e., an unbiased pricing algorithm with a demonstrated financial benefit); after the media event, the racial gap in Smart Pricing usage widened by 61.2%. A controlled experiment further demonstrates the mechanism: awareness of algorithmic racial bias differentially affects the expected financial benefits of Smart Pricing along racial lines, increasing the expected benefits of Smart Pricing among white hosts and reducing the expected benefits of Smart Pricing among Black hosts. Theoretically, this research contributes to the nascent literatures on algorithmic bias and algorithm aversion and the classic literature on judgment and decision making. Practically, it offers important implications for policy makers, firms, and media outlets that wish to prevent the public's awareness of certain algorithmic biases from spilling over to algorithms that are unbiased and beneficial.

Keywords: Algorithmic Bias, Algorithmic Racial Bias, Airbnb, Smart Pricing, Algorithm Usage, Overgeneralization

1. Introduction

Algorithmic bias, or machine bias, refers to systematic and repeatable errors in an algorithm that lead to differential treatment on the basis of race, gender, age, or another demographic factor (Fu et al. 2020a, Fu et al. 2021). While algorithms have been widely used in almost every industry for more than a decade (PwC Report, 2015), algorithmic bias went *largely* unnoticed by the public—until May 2016, when a series of news articles on algorithmic racial bias went viral and triggered a nation-wide outrage on social media.¹ Algorithms that received extensive media coverage included COMPAS (a widely used recidivism prediction algorithm, revealed to be more likely to mistakenly attribute a high risk of recidivism to low-risk Black defendants than to their low-risk white counterparts),² algorithms behind Google’s image search engine (on which a search for “three white teenagers” yielded happy images and a search for “three Black teenagers” yielded mug shots),³ and algorithms for Amazon Prime (which excluded predominantly Black zip codes from the Free Same-Day Delivery service).⁴ The articles awakened a public horror that algorithms, just like humans, can exhibit racial bias.

Not surprisingly, the specific algorithms featured in news stories received a strong backlash,⁵ but little is known about how the new awareness of racial biases in algorithms influenced consumers’ reactions to *unbiased* algorithms. Across industries, many algorithms are race-blind by design, and their neutrality has been upheld by empirical analysis (Ganju et al. 2020, Cowgill 2017, Zhang et al. 2021); in fact, algorithms can even help *reduce* racial disparities (Ganju et al. 2020; Zhang et al. 2021). Unfortunately, the assessment of algorithmic bias is difficult even for well-trained computer scientists who have access to abundant computing resources and data, let alone typical consumers (Fu et al. 2020a). An important question, then, is: Does awareness of racial bias in some algorithms deter disadvantaged consumers from adopting a different, unbiased algorithm (i.e., a spillover effect), despite the algorithm’s significant financial, education, and/or health benefits?

The current research explores this question in the context of Airbnb, an online sharing economy platform on which people can list and find lodging. In November 2015, Airbnb introduced a pricing algorithm: Smart Pricing, a free, web-embedded feature designed to help hosts optimize their rental prices

¹ Prior to 2016, a few blogs and articles wrote about how sample size disparities can lead to biased outcomes in data-driven algorithms, but general public awareness did not emerge until the May 2016 stories.

² <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

³ <https://www.washingtonpost.com/news/morning-mix/wp/2016/06/10/google-faulted-for-racial-bias-in-image-search-results-for-black-teenagers/>

⁴ <https://www.computerworld.com/article/3068622/amazon-prime-and-the-racist-algorithms.html>;

⁵ Facebook and Instagram users boycotted the two platforms over their biased algorithm for hate speech detection. Tweets written by Black users were 1.5 times more likely to be flagged as “offensive” by the algorithm than tweets written by other users. <https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter>.

for local demand dynamics (Li et al. 2016; Li and Srinivasan 2019). If a host enables Smart Pricing, then the algorithm automatically adjusts the listed nightly rate using complex machine learning models that determine the optimal price based on an extensive set of factors (Ye et al. 2018).⁶ Smart Pricing does not consider host characteristics such as race. Empirically, the algorithm has proven financially beneficial for both Black and white hosts; it offers a similar price adjustment for both racial groups, and it increases the average daily revenue by 18% among Black hosts and 8% among white hosts (Zhang et al. 2021).

To study how awareness of racial bias in some algorithms influences the usage of unbiased, beneficial algorithms, we explored a rich dataset of 8,175 unique Airbnb properties listed by either Black or white hosts in the US. We monitored Smart Pricing usage for a year (November 2015–November 2016) and analyzed usage trends before and after the extensive media coverage of algorithmic racial bias in May 2016 using the inverse probability of treatment weighting (IPTW) method and a Difference-in-Differences (DiD) analysis. Before May 2016, Black hosts were less likely to use Smart Pricing than white hosts; after the media event, the gap widened by 61.2%. We explored the reasons for the widened racial disparity in a controlled experiment in which we manipulated participants’ awareness of algorithmic racial bias and measured the expected financial benefits of using Smart Pricing. Results showed that awareness of algorithmic racial bias hurt Black participants’ expectations for the algorithm’s benefit for Black hosts and improved white participants’ expectations for the algorithm’s benefit for white hosts.

The present studies contribute to several streams of research. First, while the nascent literature on algorithmic bias has greatly advanced our understanding of bias detection (Simoiu et al. 2017), sources (Caliskan et al. 2017), and mitigations (Bechavod and Ligett 2017, Chouldechova and Roth 2018), no research, to the best of our knowledge, has studied algorithmic bias from the consumers’ perspective. The present research extends the scope of the algorithmic bias literature by investigating how awareness of algorithmic racial bias may differentially affect the behaviors of consumers of different races.

Second, recent work in the consumer behavior literature finds that people may be reluctant to rely on algorithms to make forecasts and decisions even when the algorithms clearly outperform humans. Prior research has identified several explanations: People have a lower tolerance for mistakes made by algorithms than for the same mistakes made by humans (Dietvorst et al. 2015, 2018). Also, people incorrectly believe that algorithms are less capable of accounting for consumers’ unique situations and characteristics (Longoni et al. 2020) and performing subjective tasks (Castelo et al. 2019). The current research contributes to the consumer behavior literature by discovering a novel cause of algorithm aversion: awareness of algorithmic

⁶ Airbnb does not disclose its proprietary algorithm for Smart Pricing, but the company describes some of the key inputs, including the season, hotel rates, and property characteristics. For more details: <https://blog.airbnb.com/smart-pricing>.

racial bias. While the previously documented causes suppress the use of algorithms across all racial groups equally, the present research reveals that awareness of algorithmic racial bias deters only the disadvantaged group from adopting algorithms, thereby exacerbating the already-pronounced racial gap in technology adoption.

Third, our work builds on the notion that people are insufficiently sensitive to situations and tend to overgeneralize their responses beyond situations in which the responses are valid (e.g., Arkes and Ayton 1999, Baron 2000, Hsee, Yang and Li 2019, Tversky and Kahneman 1974, Yang, Hsee, and Li 2021). Our findings enrich the literature on judgment and decision making by documenting a new form of overgeneralization in the context of technology adoption. That is, having learned that certain algorithms exhibit racial bias, consumers behave as if all algorithms are biased, so disadvantaged consumers become less likely than advantaged consumers to use algorithms that are unbiased and even beneficial.

The present research identifies a serious problem that affects consumer welfare and racial equity. Algorithms have been found to outperform human experts on a wide range of tasks and in diverse contexts (Kleinberg et al. 2018, Zhang et al. 2021, Fu et al. 2021), and the use of algorithms is usually beneficial to users. If awareness of algorithmic racial bias makes disadvantaged groups less willing than advantaged groups to use an algorithm, then the algorithm will only exacerbate the already-pronounced racial income inequality. Furthermore, a racial disparity in algorithm usage will cause disadvantaged groups to be underrepresented in the data, which may lead to more algorithmic racial bias in the future as algorithms learn the patterns that are most strongly represented in the data (Barocas and Selbst 2016; Cogwill and Tucker 2020). In Section 4, we discuss the implications of the present research for policy makers, firms, and media outlets, and we consider strategies for reducing consumers' tendency to overgeneralize knowledge about certain algorithmic biases to algorithms that are unbiased and beneficial.

2. Empirical Data and Analyses: An Airbnb Context

We examined the impact of awareness of algorithmic racial bias by analyzing trends in the usage of Smart Pricing among white and Black Airbnb hosts. Specifically, we compared usage before and after May 2016, when extensive media coverage about algorithmic racial bias attracted the public's attention—an exogenous shock to hosts' awareness of algorithmic racial bias.

2.1. Data Description

Our data pertains to 8,175 Airbnb properties in 324 zip codes in the United States over the course of the 12 months following the launch of the pricing algorithm in November 2015. White hosts ran 7,188 of the properties while Black hosts ran 987.

2.1.1. Usage of Airbnb’s Smart Pricing Algorithm

We obtained data on each host’s usage of Smart Pricing (*Algorithm Usage*) by scraping the property calendar webpages, which report whether Smart Pricing was used on each day. Specifically, the Smart Pricing algorithm is enabled (vs. disabled) on any given calendar day if the *pricing type* variable in the HTML source code appears as “demand_based_pricing” (vs. “customize” or “default”). $Algorithm_Usage_{it}$ is the proportion of days in month t on which the Smart Pricing algorithm was used for property i , scaled by 100. If $Algorithm_Usage_{it} = 0$, then property i did not use Smart Pricing in month t . If $Algorithm_Usage_{it} = 100$, then property i used Smart Pricing every day in month t .

2.1.2. Airbnb Property Characteristics

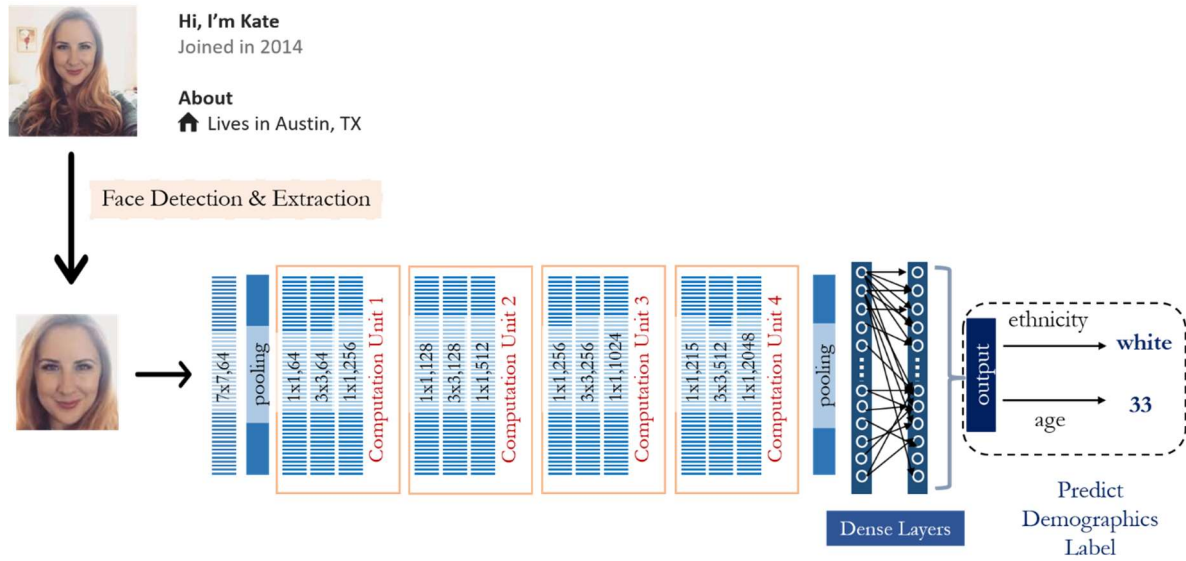
We obtained property characteristics from AirDNA, a third party that specializes in tracking and collecting Airbnb data.⁷ Characteristics included property size (e.g., number of bedrooms), amenities (e.g., fireplace, elevator, AC), location (e.g., zip code), number of accumulated reviews (*Number of Reviews*), number of property photos (*Number of Photos*), number of days in month t for which the property was booked (*# Reservation Days*), the average listing price in month t (*Nightly Rate*), and rules for guests (*Minimum Stay*, *Max Guests*, *Instant Book Enabled*, *Security Deposit*). We used the zip code to collect additional neighborhood characteristics and socioeconomic data from multiple public sources: ACS (American Community Survey), Walkscore.com, and Zillow (see Supplementary Appendix A for details).

2.1.3. Airbnb Host Characteristics

For each host, we obtained the number of Airbnb listings (*Number of Listings*) from AirDNA, and we used a deep learning classifier to predict the host’s race (*White*, *Black*, or *Other*) and age (1–100) from their profile photo. Our classifier is based on the ResNet-50, which was trained on more than 3 million face photos and achieved state-of-the-art accuracy in face recognition (Cao et al. 2018). We fine-tuned (optimized) the model on a face dataset containing more than 500,000 face photos of celebrities of a known race and age (see Figure 1). On a hold-out test set (i.e., a sample that was not used for training), the optimized model predicted race with an average accuracy of 93% and predicted age with a mean absolute error of 4.8 (see Supplementary Appendix B for technical details). We constructed the sample out of hosts who were predicted to be either Black or white.

⁷ <https://www.airdna.co/>

Figure 1 Classifying Key Demographics of Airbnb Hosts



2.2. Summary Statistics

Table 1 summarizes the key variables and statistics in our sample (see Supplementary Appendix A for a full list of variables and their statistics). The average algorithm usage was 5.09%.⁸

Table 1 Summary Statistics

VARIABLES	Mean	Std. Dev.	Median
Smart Pricing Usage			
Algorithm Usage %	5.09	20.74	0
Property Characteristics			
Number of Reservation Days	7.87	10.26	1
Nightly Rate	182.63	175.1	139
Security Deposit	160.19	325.55	0
Max Guests	3.3	2.16	2
Number of Reviews	29.62	41.15	14
Number of Photos	16.72	11.79	14

⁸ The low average usage of Smart Pricing reflects a low overall adoption rate: only about 17% of the properties ever used the pricing algorithm during our observational window.

Minimum Stay	2.74	3.15	2
Instant Book Enabled	.11	.32	0
Number of Bedrooms	1.31	.86	1

Host Characteristics

Number of Listings	2.8	6.33	1
Host Age	35.75	10.02	34.3
White (host race)	.88	.33	1
Black (host race)	.12	.33	0

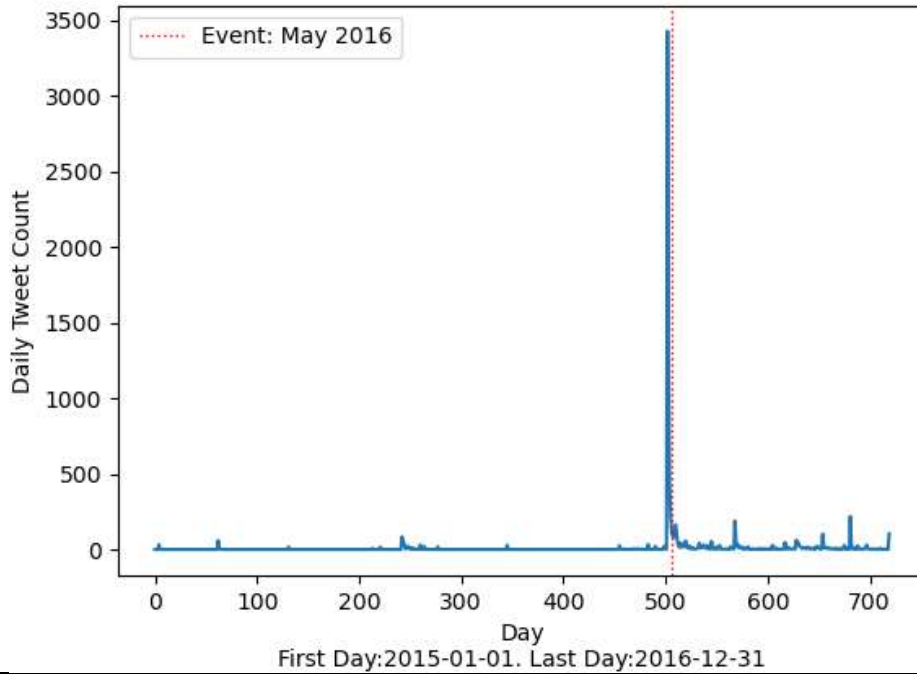
Notes: The summary statistics were computed for all properties in our sample (i.e., all those listed by Black and white hosts). The values for the two racial groups (White, Black) refer to the proportion of properties listed by hosts of the given race.

2.3. Empirical Study: Usage of Smart Pricing Before versus After the Media Coverage of Algorithmic Racial Bias

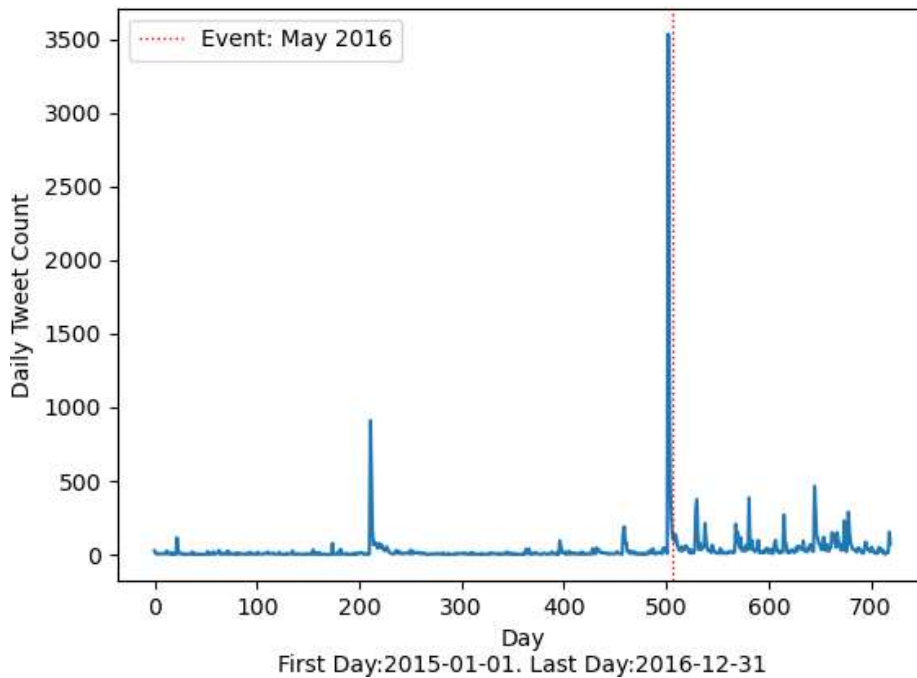
On May 23rd, 2016, ProPublica published a study (Angwin et al. 2016) on a criminal risk-assessment algorithm that assigned a higher risk score to Black individuals than to their white counterparts. The study was soon covered by major media outlets (e.g., NPR, New York Times, The Guardian) and sparked a series of articles featuring algorithmic racial bias in other contexts (Ingold and Soper 2016). To verify that the ProPublica study sparked widespread popular interest in the topic, we used Twitter API (see Section F.1.1 of the Supplementary Appendix for details) to retrieve all Tweets posted in 2015–2016 that contained “ProPublica Bias,” “Machine Bias,” or “Algorithms Biased.” In Figure 2, there is a visible spike in the use of all three terms shortly after the ProPublica study was published (indicated by the vertical red dashed line). Hence, we define the ProPublica study as *the event*, and we split our data into two periods: the *pre-event period* (November 2015–May 2016) and the *post-event period* (June 2016–November 2016).

Figure 2 Social Media Trends in Awareness of Algorithmic Bias: Daily Tweet Count Containing Queried Keywords

(a) "ProPublica Bias"
Tweet Count for Key words:Propublica Bias

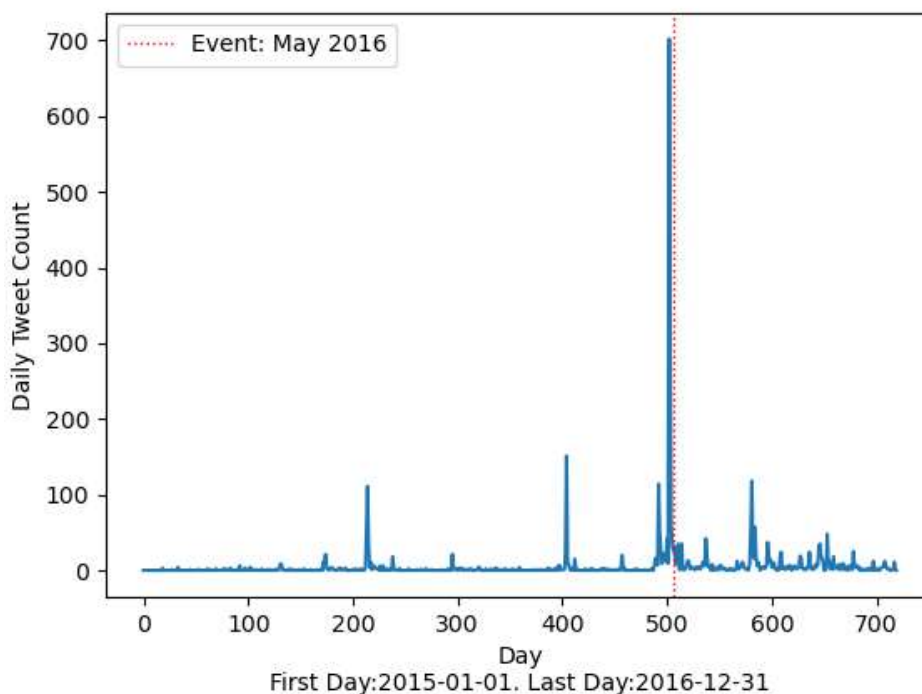


(b) "Machine Bias"
Tweet Count for Key words:Machine Bias



(c) “Algorithm Biased”

Tweet Count for Key words:Algorithm Biased



Notes: The horizontal axis indicates the days over a two-year period. The vertical axis indicates the number of Tweets posted each day. The red dashed line marks May 23rd, 2016 (when the ProPublica study was published, i.e., the event).

We computed the average usage of Smart Pricing on a monthly basis among Black hosts and white hosts and observed that the racial gap in Smart Pricing usage widened significantly following the event: Black hosts were 34.8% less likely than white hosts to use Smart Pricing in the *pre-event period* (white hosts: 1.55%, Black hosts: 1.15%)⁹ and 44.9% less likely in the *post-event period* (white hosts: 10.72%, Black hosts: 7.40%).

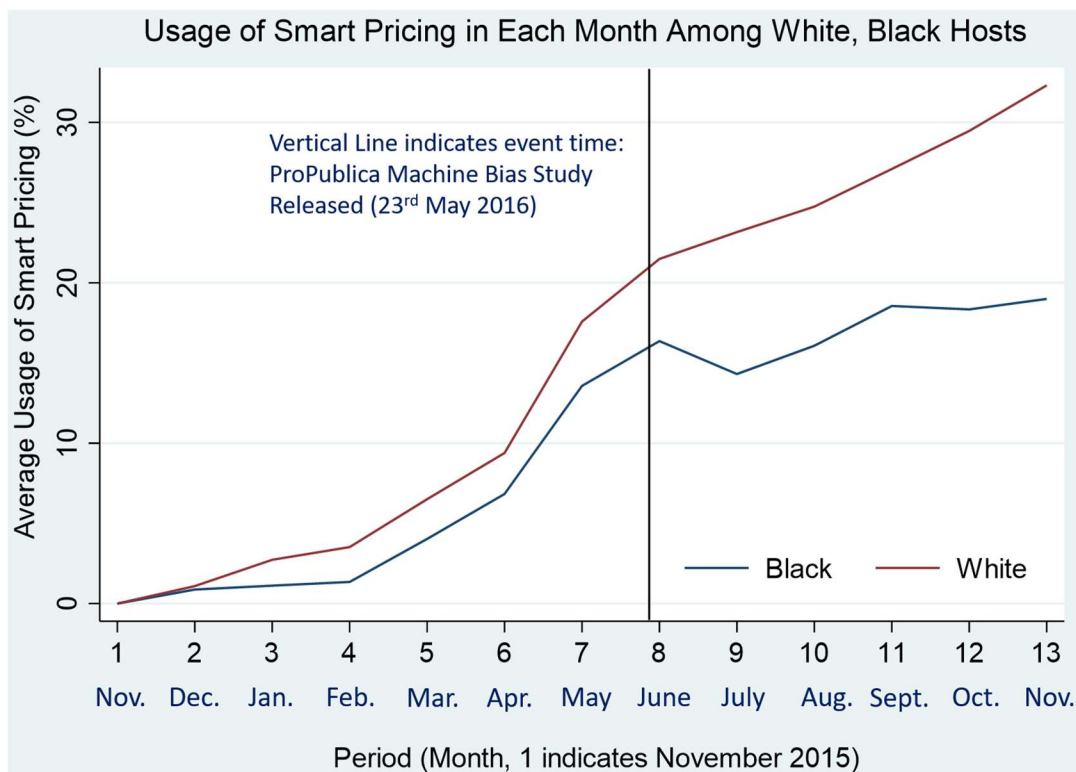
Next, to control for potential differences in host and property characteristics between Black and white hosts (Goldfarb and Prince 2008), we used the inverse probability of treatment weighting approach (IPTW, a propensity score weighting method) to construct a weighted sample with comparable groups of white and Black hosts (Bitler et al. 2006; Giorcelli 2019). The IPTW analysis first estimated a logistic regression $\sigma(\cdot)$: $Black_i \sim \sigma(X_i, \alpha)$. The dependent variable, $Black_i$, indicates whether host i is Black or not, and the independent variables, X_i , are the observed characteristics as described in Section 2.1. Using the

⁹ The racial gap in algorithm usage could plausibly be explained by differences other than race (e.g., the location of properties, socioeconomic status). These factors are likely to be time-invariant in our data window and are controlled for through property fixed effects in the DiD model.

estimated $\hat{\alpha}$, we computed an estimated propensity score: $\widehat{ps}_i = f(\hat{\alpha}X_i)$, where $\{\widehat{ps}_i\}_{i=1}^I$ is our sample weight. In Supplementary Appendix Section C.1, we validate our IPTW method by showing that the unweighted Black and white groups are not sufficiently comparable in terms of the covariates X , but the weighted groups are comparable at the standard threshold.

Figure 3 depicts the average usage of Smart Pricing among Black hosts (blue curve) and white hosts (red curve). A visual examination of the plot suggests that Smart Pricing usage was similar between the two groups *approaching* the event (vertical line) yet diverged *following* the event.¹⁰

Figure 3 Average Usage of the Smart Pricing Algorithm Among White vs. Black Hosts: An IPTW Weighted (Balanced) Sample



We then estimated a DiD regression (Equation 1) on the IPTW sample to assess how the usage of Smart Pricing changed after the event (relative to the trend established before the event):

¹⁰ Note that algorithm usage started off very low among both white and Black hosts and gradually trended upward over the course of the year, perhaps due to Airbnb’s promotion efforts and the diffusion of Smart Pricing as a new service for hosts (e.g., hosts became aware of the algorithm; hosts discussed the algorithm with each other and then decided to try it).

$$\begin{aligned}
\text{Algorithm_Usage}_{it} = & \text{Property}_i + \beta \cdot (\text{Black}_i \times \text{After_ProPublica}) \\
& + \gamma \cdot \text{Controls}_{it} + \text{Seasonality}_{it} + \varepsilon_{it}
\end{aligned}
\tag{1}$$

where *After_ProPublica* indicates the post-event period, and the coefficient β estimates the relative trend in usage among Black hosts compared to white hosts in the post-event period. *Property_i* is the property fixed effect and absorbs the time-invariant unobserved individual characteristics (including location-related factors such as the neighborhood’s infrastructure and socioeconomic characteristics). *Controls_{it}* are the time-varying host and property characteristics that may correlate with algorithm usage. For example, the current price of and demand for a property may affect the host’s decision whether to use the algorithm. We also considered the possibility of seasonal patterns in algorithm usage; *Seasonality_{it}* captures the city-month fixed effect and effectively controls for the overall time-trend in algorithm usage (i.e., that adoption increased over time as Airbnb promoted the algorithm).

As reported in Table 2, the estimated coefficient of *Black* \times *After_ProPublica* is negative and significant ($b = -4.768, p < 0.001$), suggesting that the racial gap in Smart Pricing usage widened by 61.2% from the pre-event period to the post-event period (*pre-event*, the racial gap in the algorithm usage was 7.78%; *post-event*, the gap widened to $7.78\% + 4.768\% = 12.55\%$). Critically, a validation check upheld the parallel trends assumption in the DiD model (Angrist and Pischke 2008): *pre-event* usage trendlines did not differ between Black and white hosts (see Supplementary Appendix Section C.2). That is, in the IPTW sample, the racial gap in Smart Pricing usage remained constant before the event and increased only *after* the event. We also conducted a series of robustness tests to rule out alternative mechanisms (see Supplementary Appendix Section C.3), where confounding factors such as seasonality or race-correlated variables (e.g., SES/location) plausibly could have driven the change in Smart Pricing usage.

Taken together, the empirical analyses provide robust evidence that the racial gap in Smart Pricing usage widened significantly after (vs. before) the media coverage of algorithmic racial bias. However, the underlying mechanism is less clear. The widened racial gap in Smart Pricing usage may have occurred because awareness of algorithmic racial bias increased the expected benefits of Smart Pricing among white hosts, reduced the expected benefits of Smart Pricing among Black hosts, or both. To disentangle these possibilities, we conducted a controlled lab experiment that manipulated participants’ awareness of algorithmic racial bias and measured their expectations for the financial benefits of using Smart Pricing (see Section 3).

Table 2 Individual Host’s Usage of Smart Pricing: Effects of Race and Raised Awareness of Algorithmic Racial Bias

VARIABLES	ESTIMATES	
	Coefficients	Std. Err.
After_ProPublica (months prior to May 2016 as the reference)	26.54***	(0.481)
Black × After_ProPublica (White as the reference race)	-4.768***	(0.687)
Number of Photos	0.0660*	(0.0292)
Number of Reviews	0.243***	(0.0104)
Property Nightly Rate _{t-1}	0.00314	(0.00937)
Minimum Stay	0.277***	(0.0517)
Security Deposit	0.00480***	(0.000374)
# Blocked Days _{t-1}	0.0788***	(0.0126)
# Reservation Days _{t-1}	0.274***	(0.0147)
Instant Book Enabled	6.198***	(0.410)
Max Guests	-0.331	(0.258)
Number of Listings	-0.821***	(0.111)
Seasonality Fixed Effect	City-Month Property	
Observations	102251	
R-squared	0.48	

Note: The regression was estimated on a dataset that included only the observations (*Algorithm_Usage_{it}*) where property *i* had at least one night available for booking in month *t*. IPTW sample weights were used. Heteroskedasticity-robust standard errors are in parentheses.
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

3. Randomized Experiment

3.1. Experimental Design

The experiment aimed to understand the causal relationship between raising awareness of algorithmic racial bias and racial differences in the expected financial benefits of using Smart Pricing. The experiment was pre-registered before any data collection took place (<https://aspredicted.org/blind.php?x=ii2na5>). Materials and data are publicly archived (https://osf.io/f6uad/?view_only=8e4df9eb21cd4dbbaa22f7007eec0a8a).

This experiment consisted of two parts, separated by some filler questions about demographics. Part I experimentally manipulated the awareness of algorithmic racial bias. Specifically, participants in the high-awareness condition read an article titled “Racial Bias in Algorithms” (see Appendix D), while participants in the low-awareness condition read an article titled “Protect Your Body from Injuries Caused by Digital Device Overuse” (see Appendix E). Next, participants indicated whether they had read similar

articles on this topic (1 = *not at all*, 7 = *a lot*) and whether they would share the article with others (1 = *definitely no*, 7 = *definitely yes*). These two questions gauged participants' prior exposure to similar news and intention to share the article at the time of the experiment (i.e., June 2021).

Part II measured the expected financial benefits of using Smart Pricing, which did not appear in either of the news articles in Part I. Specifically, participants learned about Airbnb and how hosts rent out properties. Participants also learned that there are two ways for hosts to set the nightly price: (1) hosts can set the nightly price themselves, or (2) hosts can enable Airbnb's pricing algorithm (named "Smart Pricing") to set the price automatically (based on a variety of factors) within the range specified by the hosts. Participants then predicted how much Smart Pricing would influence the earnings of white hosts and Black hosts by selecting percentages from -30% to +30%, inclusive.

3.2 Data and Results

We posted the study on Prolific for 400 participants (200 white and 200 Black) using Prolific's prescreening tool ("What ethnic group do you belong to?"). Three hundred ninety-seven participants completed the study for a fixed nominal payment. As pre-registered, we excluded 58 participants who failed two attention checks or reported they were neither "White" nor "Black," leaving a final sample of 339 participants (186 females, 2 other, $M_{\text{age}} = 26.47$, $SD = 8.05$) for analyses.¹¹

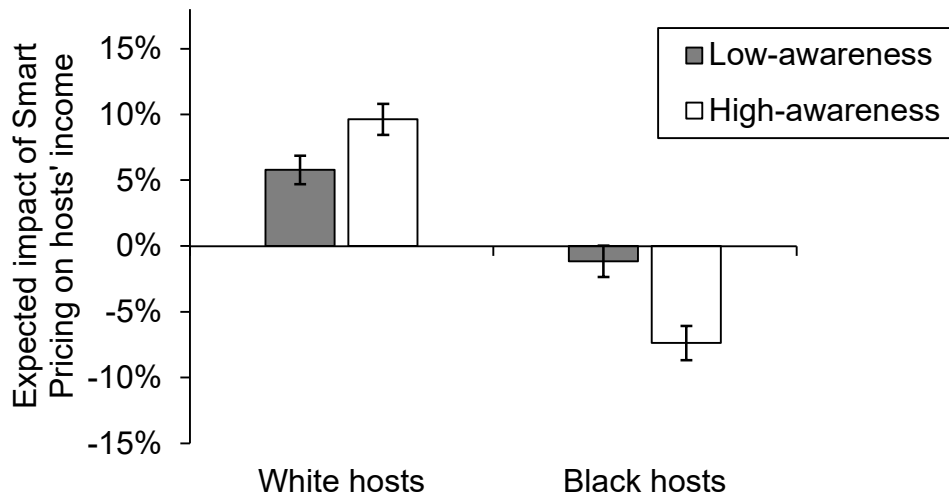
First, to test whether awareness of algorithmic racial bias influenced the expected financial outcome of using Smart Pricing for the two racial groups, we conducted two separate 2 (awareness: low vs. high; between-subjects) \times 2 (expected financial outcome: white hosts vs. Black hosts; within-subjects) mixed ANOVAs, one for white participants and one for Black participants. The analysis revealed significant awareness \times expected financial outcome interactions among both white participants ($F(1, 173) = 16.69$, $p < .001$, $\eta_p^2 = .088$) and Black participants ($F(1, 162) = 5.97$, $p = .016$, $\eta_p^2 = .036$). Specifically, awareness of algorithmic racial bias significantly widened the expected outcome gap between white hosts and Black hosts regardless of the participant's own race, but the nature of the expected outcome gap differed between white and Black participants: Among white participants, awareness of algorithmic racial bias increased the expected benefit of Smart Pricing for white hosts ($M_{\text{low-awareness}} = 5.7\%$, $SD = 10.9\%$ vs. $M_{\text{high-awareness}} = 9.6\%$, $SD = 10.2\%$; $F(1, 173) = 5.70$, $p = .018$, $\eta_p^2 = .032$) and increased the expected harm of Smart Pricing for Black hosts ($M_{\text{low-awareness}} = -1.2\%$, $SD = 11.7\%$ vs. $M_{\text{high-awareness}} = -7.4\%$, $SD = 11.6\%$; $F(1, 173) = 12.35$, $p = .001$, $\eta_p^2 = .067$; see Figure 4, top panel). Among Black participants, however, the expected benefit of Smart Pricing for white hosts was similar across conditions ($M_{\text{low-awareness}} = 14.3\%$, $SD = 11.9\%$ vs. $M_{\text{high-awareness}} = 14.8\%$, $SD = 13.7\%$; $F(1, 162) = .065$, $p = .799$, $\eta_p^2 = 0$), while only participants in the high-awareness condition expected that Black hosts would lose income by using Smart Pricing ($M_{\text{low-awareness}} =$

¹¹ 11.2% of white participants and 15.9% of Black participants failed the attention checks ($\chi^2(1) = 1.88$, $p = .17$).

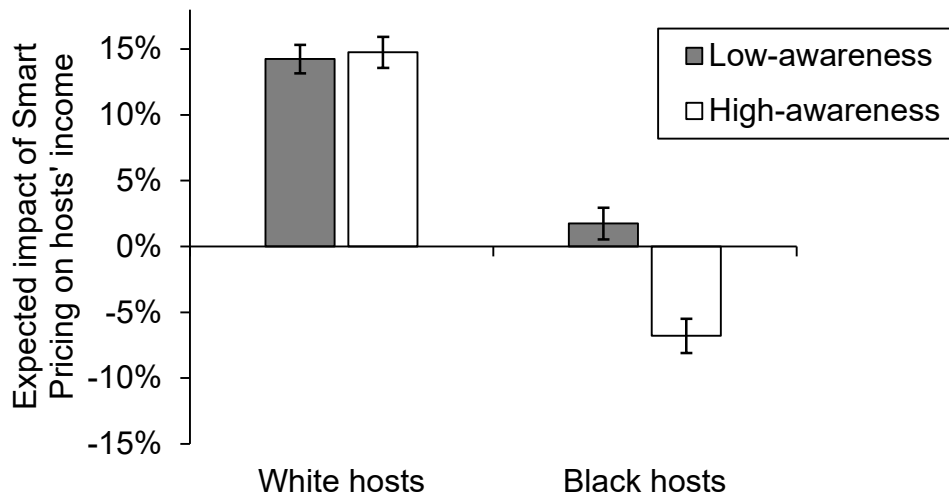
1.75%, SD = 16.2% vs. $M_{\text{high-awareness}} = -6.8\%$, SD = 17.1%; $F(1, 162) = 10.75$, $p = .001$, $\eta_p^2 = .062$; see Figure 4, bottom panel).

Figure 4 Expected Financial Outcome of Smart Pricing for White versus Black Hosts, as Predicted by White versus Black Participants

White Participants



Black Participants



In addition, we examined participants' prior exposure to similar news and intention to share the news article with others by conducting 2 (awareness: low vs. high) \times 2 (participant's race: white vs. black) between-subjects ANOVAs. Results showed that while white participants had more prior exposure to news regarding injuries caused by digital device overuse than Black participants ($M_{\text{white}} = 3.47$, $SD = 1.75$ vs. $M_{\text{Black}} = 2.46$, $SD = 1.89$; $F(1, 335) = 12.34$, $p = .001$, $\eta_p^2 = .036$), they did not differ in their exposure to news on algorithmic racial bias ($M_{\text{white}} = 2.86$, $SD = 1.89$ vs. $M_{\text{Black}} = 2.92$, $SD = 2.06$; $F(1, 335) = .033$, $p = .855$, $\eta_p^2 = 0$). Moreover, in both awareness conditions, Black participants reported a higher intention to share the news article than white participants (for algorithmic racial bias: $M_{\text{white}} = 5.10$, $SD = 1.62$ vs. $M_{\text{Black}} = 5.80$, $SD = 1.63$; $F(1, 335) = 6.44$, $p = .012$, $\eta_p^2 = .019$; for device overuse: $M_{\text{white}} = 4.05$, $SD = 1.97$ vs. $M_{\text{Black}} = 5.60$, $SD = 1.76$; $F(1, 335) = 33.60$, $p < .001$, $\eta_p^2 = .091$).

In sum, this experiment showed that raising awareness of algorithmic racial bias led both Black and white participants to expect that Smart Pricing—an unbiased algorithm—would hurt Black hosts, and raising awareness also led white participants to expect a greater benefit for white hosts.¹² This result suggests that the widened racial gap in Smart Pricing usage *post-event* (observed in Section 2) may have been driven by both an increased expected benefit of Smart Pricing among white hosts and a reduced expected benefit of Smart Pricing among Black hosts.

4. General Discussion

The current research is the first attempt to understand how consumers' awareness of the existence of algorithmic bias in one context influences their receptivity to an unbiased, beneficial algorithm in a different context. Evidence from an Airbnb dataset and a controlled experiment suggests that awareness of algorithmic racial bias in one context leads Black consumers to anticipate harm from an unrelated (and unbiased) algorithm, while the same knowledge leads white consumers to anticipate a greater benefit. Diverging expectations result in a greater racial disparity in the rate of algorithm adoption and, subsequently, in the actual benefits reaped from the technology.

Theoretically, our work significantly extends the nascent literatures on algorithmic bias and algorithm aversion as well as the classic literature on judgment and decision making. Specifically, while the prior research on algorithmic bias focuses on technical aspects (e.g., sources, solutions), we focus on consumers' perspectives. While recent work has uncovered reasons for algorithm aversion that exert a

¹² Of note, even participants who did not read the article on algorithmic racial bias (i.e., the low-awareness condition) predicted that Smart Pricing would benefit Black hosts less than white hosts. The gap is not surprising given that (1) some participants might have heard of algorithmic bias from other sources before participating in our experiment, and (2) the presence of separate questions regarding white hosts and Black hosts may have activated the concept of discrimination and therefore raised suspicion about the fairness of Smart Pricing.

similar influence on consumers of all demographics, we identify a reason—awareness of algorithmic racial bias—that differentially affects consumers of different races. Our novel example of overgeneralization in the context of a new technology adds to decades of research on judgment and decision making in traditional settings.

Our findings suggest that awareness of algorithmic bias has unintended and profound social consequences. First, it deters disadvantaged groups from adopting new technologies and enjoying the benefits of those technologies—so, ironically, awareness of algorithmic bias exacerbates existing inequalities in technology usage and income. Second, disparate rates of algorithm usage lead to imbalanced data (i.e., advantaged groups will generate more data pertaining to algorithm usage than disadvantaged groups), and this imbalance may cause the algorithms to prioritize learning the behavior of the advantaged groups over the disadvantaged groups, leading to algorithmic bias—even if there was none in the first place (Barocas and Selbst 2016; Cowgill and Tucker 2020). In other words, the avoidance of algorithms by disadvantaged groups could lead to algorithmic bias, which could further deter these groups from using algorithms, creating a vicious cycle.

Our work underscores the need for policy makers and firms (e.g., Airbnb, LinkedIn) to mitigate fears about algorithmic bias among potential algorithm users. Given the importance and difficulty of detecting algorithmic bias, policy makers should encourage firms to disclose fairness-related information about their algorithms, such as the prediction accuracy across demographic groups, the variables that the algorithms use to capture any demographic differences, why such variables should not generate discriminatory outcomes, and what the firms have done to minimize algorithmic bias. Transparent practices (Lakkaraju et al. 2017) may restore trust and increase algorithm usage especially among disadvantaged groups (Buell et al. 2017). If firms enable consumers to discern whether an algorithm is biased, then potential algorithm users from disadvantaged groups may be less likely to overgeneralize concerns about algorithmic bias to algorithms that are unbiased and beneficial.

Our research also yields implications for media outlets that wish to inform the public about algorithmic bias. Media outlets have played (and will continue to play) a critical role in raising awareness of algorithmic bias. Our research suggests that the media should approach this topic carefully, as consumers who are exposed to the notion of biased algorithms may incorrectly perceive that all (or most) algorithms are biased. A perception of widespread algorithmic bias can lead to a series of negative consequences that are detached from the presence of actual bias. More accurate and responsible reporting on algorithmic bias would provide examples of both biased and unbiased algorithms, thereby encouraging consumers to make more nuanced judgments about algorithms rather than overgeneralizing one upsetting finding from a sensational news article. For example, if a media outlet wished to discuss algorithmic racial bias in

healthcare, the outlet could present the findings of racial bias documented by Obermeyer et al. (2019) alongside the counterexample documented by Ganju et al. (2020).

Finally, our research offers inspiration for future research at the intersection of equity and technology adoption. While our research is focused on how awareness of algorithmic bias influences algorithm adoption rates along racial lines, future research may investigate whether the effect generalizes to other types of algorithmic biases involving gender or age. Moreover, future research may investigate a potential spillover effect between disadvantaged groups (i.e., perhaps members of one disadvantaged group may be reluctant to use algorithms after learning about algorithmic bias toward a different disadvantaged group). We believe that research addressing these important and pressing questions will have significant impacts.

5. References

- Angrist JD, Pischke JS (2008) *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton University Press).
- Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica* (May 23), <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- Arkes HR., Ayton P (1999) The sunk cost and Concorde effects: Are humans less rational than lower animals? *Psychol Bull.* 125(5): 591–600.
- Baron J (2000) The effects of overgeneralization on public policy. Working paper. University of Pennsylvania, Pennsylvania.
- Barocas S, Selbst AD (2016) Big data's disparate impact. *Calif. L. Rev.* 104(4): 671-732.
- Bechavod Y, Ligett K (2017) Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*.
- Bitler MP, Gelbach, JB, Hoynes HW (2006) What mean impacts miss: Distributional effects of welfare reform experiments. *Am Econ Rev.* 96(4): 988-1012.
- Buell RW, Kim T, Tsay CJ (2017) Creating reciprocal value through operational transparency. *Manag Sci.* 63(6): 1673-1695.

- Caliskan, A, Bryson J, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334): 183-186.
- Castelo N, Bos MW, Lehmann DR (2019) Task-dependent algorithm aversion. *J Mark Res.* 56(5): 809-825.
- Chouldechova A, Roth A (2018) The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*.
- Cowgill B (2017) Automating judgement and decision-making: Theory and evidence from résumé screening. Working paper: <https://ssrn.com/abstract=3361280>
- Cowgill B, Tucker CE (2020) Algorithmic fairness and economics. *J Econ Perspect.* in press.
- Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Experiment. Psych.: General* 144(1):114–126.
- _____ (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Manag Sci.* 64(3):1155–1170.
- Goldfarb A, Prince J (2008) Internet adoption and usage patterns are different: Implications for the digital divide. *Inform. Econom. Policy* 20(1):2–15.
- Ganju KK, Atasoy H, McCullough J, Greenwood B (2020) The role of decision support systems in attenuating racial biases in healthcare delivery. *Manag Sci.* 66(11): 5171-5181.
- Fu R, Huang Y, Singh PV (2020a) Artificial Intelligence and Algorithmic Bias: Source, Detection, Mitigation, and Implications. *INFORMS Tutorials in Operations Research:* 39-63. <https://pubsonline.informs.org/doi/abs/10.1287/educ.2020.0215>
- _____ (2021) Crowds, lending, machine, and bias. *Inf Syst Res.* forthcoming.
- Fu R, Aseri M, Singh PV, Srinivasan K (2021) ‘Un’fair machine learning algorithms. Forthcoming at *Management Science:* <https://ssrn.com/abstract=3408275>.
- Giorelli M (2019) The long-term effects of management and technology transfers. *Am Econ Rev.* 109(1): 121-152.
- Hsee CK, Yang Y, Li X (2019) Relevance insensitivity: A new look at some old biases. *Organ Behav Hum Decis Process* 153, 13-26.
- Ingold D, Soper S (2016) Amazon doesn’t consider the race of its customers. Should it? *Bloomberg* (April 21), <https://www.bloomberg.com/graphics/2016-amazon-same-day/>

- Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan J, (2018) Human decisions and machine predictions. *Q J Econ.* 133(1): 237–293.
- Lakkaraju H, Kamar E, Caruana R, Leskovec J (2017) Interpretable and explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*.
- Li J, Moreno A, Zhang DJ (2016) Pros vs Joes: Agent pricing behavior in the sharing economy. Working Paper, University of Michigan, Ann Arbor.
- Li H, Srinivasan K (2019) Competitive dynamics in the sharing economy: An analysis in the context of Airbnb and hotels. *Marketing Sci.* 38(3):365–391.
- Longoni C, Bonezzi A, Morewedge CK (2019) Resistance to medical artificial intelligence. *J Consum Res.* 46(4): 629-650.
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453.
- PwC Report (2015) Sizing the prize, <https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf>
- Simoiu C, Corbett-Davies S, Goel S (2017) The problem of infra-marginality in outcome tests for discrimination. *Ann. Appl Stat.* 11(3): 1193-1216.
- Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185(4157): 1124–1131.
- Zhang, Shunyuan and Mehta, Nitin and Singh, Param Vir and Srinivasan, Kannan (2021) Can an AI Algorithm Mitigate Racial Economic Inequality? An Analysis in the Context of Airbnb. Forthcoming at *Marketing Science*: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3770371.
- Yang Y, Hsee C, Xilin L (2021) Prediction biases: An integrative review, *Curr Dir Psychol Sci.* 30(3): 195-201.
- Ye, P., Qian, J., Chen, J., Wu, C.H., Zhou, Y., Mars, S.D., Yang, F. and L. Zhang (2018), “Customized Regression Model for Airbnb Dynamic Pricing,” KDD 18, 19–23.