A Lambrecht and C Tucker
Apparent algorithmic discrimination and real-time algorithmic learning in digital search advertising
Article

Lambrecht, A and Tucker, C

(2024)

Apparent algorithmic discrimination and real-time algorithmic learning in digital search advertising.

Quantitative Marketing and Economics.

ISSN 1570-7156

(In Press)

DOI: https://doi.org/10.1007/s11129-024-09286-z

Springer Verlag (Germany)
https://link.springer.com/article/10.1007/s11129-0...

Apparent Algorithmic Discrimination and Real-Time Algorithmic Learning in Digital Search Advertising

Anja Lambrecht and Catherine Tucker*

May 22, 2024

Statements and Declarations: This research was not funded by any companies or any external grant. The authors have no competing interests to declare that are relevant to the content of this article. The authors have no financial or proprietary interests in any material discussed in this article. However, both authors have consulted widely outside of this research. Catherine Tucker's conflict of interest statement may be found at `https://mitmgmtfaculty.mit.edu/cetucker/disclosure/`. Anja Lambrecht's disclosure statement may be found at `https://www.london.edu/faculty-and-research/faculty-profiles/l/lambrecht-a`.

*Anja Lambrecht is Professor of Marketing at London Business School, alambrecht@london.edu. Catherine Tucker is the Sloan Distinguished Professor of Marketing at MIT Sloan School of Management, Cambridge, MA and Research Associate at the NBER, cetucker@mit.edu, Thank you to NSF CAREER Award 6923256 for financial support. All errors are our own. We thank for their helpful comments: Garrett Johnson, Thomas Otter, Caroline Wiertz, Hema Yoganarasimhan and Xu Zhang; seminar participants at the eQSM virtual seminar, National University Singapore, IDC Herzliya, Rotterdam School of Management, the University of Chicago and the VIDE (Virtual Digital Economics Seminar); and participants at the 2020 Marketing Science conference. We are grateful to Chaoran Liu for excellent research support.

Apparent Algorithmic Discrimination and Real-Time Algorithmic Learning in Search Advertising

Abstract

Digital algorithms try to display content that engages consumers. To do this, algorithms need to overcome a 'cold-start problem' by swiftly learning whether content engages users. This requires feedback from users. The algorithm targets segments of users. However, if there are fewer individuals in a targeted segment of users, simply because this group is rarer in the population, this could lead to uneven outcomes for minority relative to majority groups. This is because individuals in a minority segment are proportionately more likely to be test subjects for experimental content that may ultimately be rejected by the platform. We explore in the context of ads that are displayed following searches on Google whether this is indeed the case. Previous research has documented that searches for names associated in a US context with Black people on search engines were more likely to return ads that highlighted the need for a criminal background check than was the case for searches for white people. We implement search advertising campaigns that target ads to searches for Black and white names. Our ads are indeed more likely to be displayed following a search for a Black name, even though the likelihood of clicking was similar. Since Black names are less common, the algorithm learns about the quality of the underlying ad more slowly. As a result, an ad is more likely to persist for searches next to Black names than next to white names. Proportionally more Black name searches are likely to have a low-quality ad shown next to them, even though eventually the ad will be rejected. A second study where ads are placed following searches for terms related to religious discrimination confirms this empirical pattern. Our results suggest that as a practical matter, real-time algorithmic learning can lead minority segments to be more likely to see content that will ultimately be rejected by the algorithm.

Keywords: Algorithmic Fairness, Algorithmic Discrimination, Advertising

1 Introduction

Algorithms are often optimized to try to ensure that consumers see content or ads they are likely to be interested in. To do this, algorithms need to use data to evaluate consumers' responses to content or ads and resolve what is often called the 'cold start' problem. For an advertising campaign, this means that if the campaign is run in parallel across multiple segments, there will be differential consequences for individuals across the segments that depend on how quickly the cold start problem is resolved. As a result, members of minority segments will be more likely to see content that will ultimately be rejected by the algorithm. This is harmful in view of a legal literature that has highlighted that concerns of algorithmic fairness apply precisely when minority groups are proportionally more likely to have a different experience than the majority segment (Hellman 2020).[1]

For an individual who is in a segment with many other people, the data to resolve the cold start problem will be provided swiftly - most likely by others - and in expectation the burden on this individual is small. However, for an individual who belongs to a segment with a small population, the data will be provided more gradually, and that person is likely going to be called upon to see content or have content associated with them which may ultimately be rejected. Therefore, minority groups may be more likely to see content that is unappealing or low quality, relative to majority groups. This paper examines this theoretical possibility, and evaluates the extent to which we observe it mattering empirically. This is important because in general, measuring the performance of algorithms is challenging, and platforms do not necessarily have incentives to do it.

We take an experimental approach in the context of Google paid search campaigns. Prior work by Sweeney (2013) documented from a user's perspective a disconcerting pattern, whereby searches for a Black name are more likely to lead to ads that suggest that person warrants a background check than searches for a white name, even though the names used in the study were purposely similar, and the implication of a need for a background check has professional consequences. By contrast, what is novel in our study we collect data from the advertiser perspective by running

[1]Throughout, the discussion in Hellman (2020) emphasizes that when considering algorithmic fairness, one should worry about the proportion – not the absolute number – of individuals that may be disadvantaged in a majority or minority group.

experimental ad campaigns which allows us access to more data on both outcomes and the workings of the algorithm. This allows us to empirically document whether the process by which advertising algorithms determine in real time if an ad is of interest to a consumer, affects the display of ads, such that searches associated with minority groups are more likely to lead to seeing unappealing content. This is because algorithmic learning requires a minimum number of observations to evaluate user response to ad content, but observations for minority groups are contributed more sparsely and at a lower speed than for the majority group, slowing the algorithm's learning about the minority group.

We conducted a search advertising campaign on Google that targeted 865 combinations of first and last names that are used either predominantly by Black or white populations in the US. We then extracted the data made available by Google to advertisers. This data collection approach has advantages over automating web scraping of search results, as Google shares with advertisers metrics that affect the placement of their ads. After six weeks, a cross-sectional analysis of our data revealed that significantly more ads were being shown next to searches for Black names than white names. We first explored whether differences in the likelihood of clicking on ads following Black-name than white-name searches could explain our results. We found that the likelihood of clicking is virtually identical. Instead, we show that because Black names are less frequently searched, as a result of individual Black names being less frequent in the population, the algorithm takes longer, on average, to learn about user preferences for the ad when a campaign targets a Black-name search than when it targets a white-name search. When the platform has learned about the ad, a process that occurs significantly more often for white-name searches, the platform tends to judge the campaign as being of low quality and, as a result, is unlikely to display it in the future. As a result, a person searching for a more uncommon search term is likely to see different ads, even in situations where the advertiser had no discernible discriminatory intent.

We then confirm this pattern in a second campaign using the context of religious affiliation. Here, we examine ads for different types of religious employment discrimination and find that ads persist for longer when they are targeted towards groups that are less searched for. We also find that this algorithmic learning process starts after a relatively small number of impressions. This mechanism of algorithmic learning has implications for the specific context of online advertising, as well as for the broader use of algorithms to parse content in real time. As data availability matters, an algorithm will learn at a lower speed for a smaller or minority group, leading to a differential quality of decisions or recommendations across groups. When taking a snapshot in time, we show empirically that this mechanism can lead to uneven outcomes, without any economic actor intending to discriminate against the minority group.

A natural question to ask is if an algorithm takes a specific number of data points to learn about the quality of content, whether learning at different speeds matters if ultimately the algorithm shows the same amount of undesirable content to a minority group and a majority group. We argue it does matter. Say we target a segment that consists of 300 Black people and a segment that consists of 3000 white people. The algorithm needs 100 observations of people from each segment engaging (or not engaging) with a piece of content to learn that a certain piece of content is undesirable. This means that in total 100 Black people and 100 white people will see the potentially objectionable content. Some might argue that seems unremarkable. However, we are arguing that it matters that a Black person is likely to be exposed to the undesirable content 33% of the time, while a white person is exposed to this content 3% of the time. This view aligns with the legal literature, that emphasizes that the likelihood of a person of a minority group experiencing something different from a member of the majority group indeed matters (Hellman 2020, Nachbar 2020, Abu-Elyounes 2020). This literature stresses that predictions made by an algorithm should be equally accurate for members of protected groups, relative to other groups, and further emphasizes that any measures need to focus on probabilities in cross-sectional outcomes. The fact we document this is not occurring due to the cold-start problem is therefore especially important.

We emphasize that though we show results for contexts where search terms are associated with race and religion, our results also apply to other paid search contexts such as when paid search algorithms are trying to determine which products to highlight to consumers. Suppose there was a paid search ad for a currency exchange which ultimately struck customers as untrustworthy and so the algorithm is likely to learn not to show it. A consumer searching for a USD:EUR exchange is far less likely to be exposed to such an ad, simply because many other consumers are likely to have conducted the same search previously, then a consumer searching for an exchange rate of NAD:MNT, or Namibian dollars to Mongolian tughriks. Such users in the minority group looking for seldom-used products will not benefit from the presence of other users like them to weed out undesirable content.

While our empirical studies focus on two distinct empirical contexts in paid search, our results are relevant beyond paid search advertising and generalize to other contexts of advertising where harm can arise from different speeds of algorithmic learning across groups. For example, a finance company offering loans at particularly high rates may target display ad campaigns by county, using the programmatic ecosystem to identify whether someone is located in a certain county. Let us assume that the algorithm, after resolving its cold-start problem, would be unlikely to show such ads as they proved unpopular. However, this learning proceeds at different speeds for urban counties that have millions of residents, such as Los Angeles County, relative to rural counties that are very sparsely populated, such as Blaine County in Nebraska, which has 470 residents. As a result, people in rural and sparsely populated counties are far more likely to view content that is subject to the cold-start problem as there are few other individuals in this segment that the algorithm could learn from. In this context, the process of algorithmic learning may therefore lead to predatory ads being a lot more likely to be displayed to rural users. In general, the extent to which this is likely to be harmful depends on the vulnerability or degree of historic disadvantage of the minority group relative to the majority, and also the degree to which the content is problematic for more vulnerable or historically disadvantaged populations.

Our paper has implications for advertisers. In the paid search context, it may not be clear to advertisers that it is possible for their ads to operate in a discriminatory fashion. After all, this is a context where the advertiser chooses which search term to advertise next to in a uniform way. We empirically document that even in this setting, algorithmic learning can lead to different outcomes for those in a majority segment compared to those in a minority segment. As a result, advertisers need to be aware that even if all ad campaigns are set up equivalently, algorithmic learning may imply that the likelihood of a user seeing them may not be identical across groups.

Since the cold start problem for algorithmic learning is a reflection of the natural operation of machine learning, little has been done to tackle this issue by platforms. Though platforms have taken steps such as stopping advertisers from using protected class data (such as gender, race, and age) for ad targeting for products such as housing and credit and employment opportunities, they have not taken similar action to ensure that the way that machine learning operates does not have differential implications across these protected classes. We hope that by emphasizing this potential for uneven outcomes, we will encourage platforms to evaluate the extent to which this happens and provide guardrails and options for advertisers who want to avoid such outcomes. Similar shifts have been achieved in recruiting practices by pharmaceutical companies for pharmaceutical trials to try and actively recruit more members of minority groups, as a result of an academic literature on how sparse data about minorities can lead to reductions in pharmaceutical efficacy for those groups (Burroughs et al. 2002). The key point for platforms to determine is whether having uniform requirements for the amount of data needed to resolve a cold-start problem itself can lead to uneven outcomes. Further, platforms can consider the extent to which algorithms should use insights across different groups targeted by the same or very similar content.

Our work has implications for policy surrounding algorithmic fairness. As far as we are aware, ours is the first paper to show the empirical importance of the process of algorithmic learning in the outcomes for minority relative to majority groups. Our empirical results suggest that the cold start problem and the learning process mean that any undesirable content will be shown to a larger share of a minority group than a majority group before the algorithm determines it to be undesirable. This means that while a single member of a majority group is unlikely to be exposed to undesirable content, a single member of a minority group is more likely to be exposed to undesirable content. We believe ours is the first paper to demonstrate the empirical importance of the process of algorithmic learning in the outcomes for minority relative to majority groups. Our empirical results also apply to digital content that might be desirable rather than undesirable. One such example of desirable content is search ads for a website that explains how to deal with employment discrimination in the workplace. We show in our second study that an individual from a group of users who is more likely to search for such information over time becomes less likely to be exposed to a helpful ad than an individual from a group of users who is less likely to be discriminated against."

1.1 Literature Review

Our paper adds to three streams of the academic literature.

First, our paper builds on a literature examining questions of algorithmic fairness in advertising. Datta et al. (2015) found that women were less likely to see ads for an executive coaching service in India, but did not determine the mechanism behind this outcome. Ali et al. (2019) and Ali et al. (2021) found that in some contexts, the landing page and, more strongly, the creative used in a campaign can affect the demographic groups to which the platform is likely to direct an ad. By contrast, Lambrecht and Tucker (2019) showed that a cost-minimizing algorithm displayed STEM career ads more to men than to women because male eyeballs are cheaper. Our research emphasizes that even in a setting where the advertiser has control, the simple mechanics inherent in the learning phase of an algorithm can still inadvertently contribute to uneven outcomes. Importantly, one difference between our results and Lambrecht and Tucker (2019) is that the mechanics of algorithmic learning apply broadly even when costs may not different across different target segments.

Second, our paper contributes to a broader debate on algorithmic fairness, including prior research in statistics (Mitchell et al. 2021), computer science (Barocas et al. 2017) and law (Hellman 2020). The empirical focus of this debate has been on algorithms assessing the risk of recidivism (Dressel and Farid 2018, Kleinberg et al. 2015, Cowgill 2018), screening resumes (Dastin 2018, Cowgill 2017), and supporting health care decisions (Obermeyer et al. 2019). This prior work emphasizes that uneven outcomes can be caused by biases in training data, either because the data collected is unrepresentative, or because it reflects existing prejudices or measurement error. In the context of the well-known issue of sample size disparity when programmers train algorithms, the worry is that the static data sets used by programmers may not contain enough data points for each group to allow an algorithm to make fairer decisions once it has been trained (Barocas et al. 2017). By contrast, in this research we discuss how the process by which algorithms learn in real time may systematically reinforce social inequity, even without discriminatory intent.

Third, our paper contributes to a literature on marketing that focuses on best management practices towards the deployment of algorithms. Some work tries to help advertisers best place bids for advertising (Tunuguntla and Hoban 2021), while other work such as Srinivasan and Sarial-Abi (2021) examines a firm's best responses when there are algorithmic failures. Ukanwa and Rust (2021) use agent-based modeling to show that in the short run, discriminatory algorithms can increase profits, and that therefore careful management and increased measurement is needed to ensure both long-term profits and societal well-being. Our paper contributes to this literature by empirically showing the importance of algorithmic distortions due to algorithmic learning, even in a setting where managers have apparent control.

2 Data Collection and Analysis

2.1 Collection of Search Advertising Data

Sweeney (2013) documented how someone who searches on Google for a name typically given to a Black person is more likely to see ads for background check services and criminal records checks than if they are searching for a name typically given to someone White. As such, ads have the potential to worsen current discrimination in hiring decisions (Bertrand and Mullainathan 2004). However, while Sweeney (2013) documented robust evidence across a large number of names, their focus was not on pinning down the mechanism which lead to these uneven outcomes. Sweeney (2013) concluded their study with "Why is this discrimination occurring? Is Instant Checkmate, Google, or society to blame? We don't yet know, but navigating the terrain requires further information about the inner workings of Google AdSense."

Our first study therefore focuses on name searches of individuals online. Name searches are important. Employers frequently search for the names of job applicants online. A recent study suggests that around 69% of employers use online search engines such as Google, Yahoo and Bing to research candidates.[2] The outcome of online searches may affect whether or not an applicant is invited to an interview and ultimately receives a job offer (Acquisti and Fong 2020). In addition, people search for names for reasons such as to learn about professional service providers, new work colleagues or potential dates, the results of which will influence whether a professional service provider such as a lawyer is hired, the attitudes of coworkers and professional progress, or the

[2] See `https://www.monster.com/career-advice/article/hr-googling-job-applicants`, `https://www.careerattraction.com/how-to-survive-being-googled-by-potential-employers/`

likelihood of finding a life partner.[3]

We implement a set of advertising campaigns from the perspective of an advertiser, with the objective of understanding what drives uneven outcomes when targeting ads to Black and white name searches. Collecting data through advertising campaigns has the advantage, relative to data-scraping methods, that we can access detailed data which Google releases about the performance of ads and that thus may inform which factors drive imbalances in the display of ads. We do this in the context of paid search advertising, an area of marketing that has been much studied in the marketing literature (Edelman et al. 2007, Ghose and Yang 2009, Rutz and Bucklin 2011).

We generate a list of names which serve as keywords for campaigns to target. Specifically, we use Sweeney (2013)'s list of first names along with the indicator of whether a name was typically given to Black or white people. This list builds on work by Fryer Jr and Levitt (2004) and Bertrand and Mullainathan (2004), which in turn were based on patterns of the census.[4] For example, while "Emily" signified that this person was likely to be a white woman, the name "Tyrone" suggested that this person was likely to be a Black man. In total, this gave us 62 first names, which we list in Table A1 in the Appendix.

Sweeney (2013) does not report the last names used in the analysis. Therefore, to collect data on last names we turned to the 2010 census.[5] We focused on most common last names in the US.[6] We combined these 14 last names with the 62 first names, resulting in 868 combinations. We emphasize that we use full names as we wish to target typical name searches for individuals, such as might occur when recruiters search for names of applicants. We determine whether a name is likely to be associated with a Black or white person exclusively based on the first name. This is based on the fact that first names tend to be closely linked to race, whereas many last names are common among both the Black and white population.

In our analysis, we preemptively dropped three of these combinations as these were the names of

[3] `https://edition.cnn.com/2011/12/07/tech/social-media/netiquette-google-stalking/index.html`, `https://edition.cnn.com/2011/12/14/tech/web/netiquette-readers-googling/index.html`

[4] Sweeney (2013) also added Latanya and Latisha to the list based on observational data.

[5] `https://www.census.gov/topics/population/genealogy/data/2010_surnames.html`

[6] We started with 20 names and dropped any last names which were over 90% Hispanic in origin to avoid drawing in names most characteristic of another minority group. Table A2 in the appendix documents this. This procedure left us with 14 individual last names.

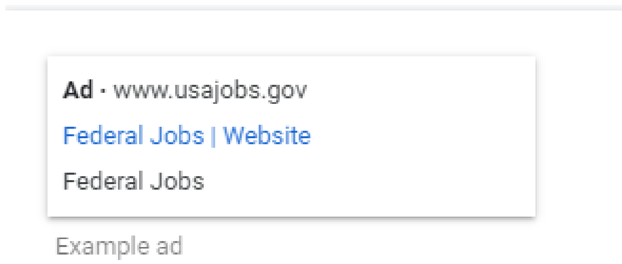

Figure 1: Ad creative

well-known individuals.[7] Including such names in our search advertising campaigns would produce a different pattern of behavior from that of a name-search for a regular person.

Using a Google advertising account, we set up 865 search advertising campaigns. Each campaign targeted one firstname-lastname combination. Each campaign instructed Google to display our ad whenever a user searched for one of the full names on our list. All campaigns used the same ad creative and text advertising information on jobs in the federal government. The format of search ads limits the information that can be displayed in an ad, and we intentionally kept the ad creative simple to ensure even interpretation. Figure 1 shows the ad creative used. All campaigns linked to the same landing page giving information on pathways into government jobs. Since the ad creative and landing page were identical across campaigns, they should not directly lead to differential inferences about the campaigns (Ali et al. 2019).

In setting up our ad campaigns, we were careful to avoid any potential spillovers between campaigns. Therefore, we set up a separate ad campaign for each keyword. Across all campaigns, we set a maximum cost per click of $10 but did not set a lifetime budget.

When a user searched for a name targeted by one of our campaigns, our ad would enter a search advertising auction, along with ads by other advertisers targeting this search term. The search advertising auction then determined whether or not our ad was displayed.

We ran all campaigns concurrently over a six-week time period in 2019. After this period, we downloaded from Google AdWords the data that is available to advertisers. Table 1 summarizes descriptives for the data. Half of the names in our data are typically associated with Black people.

[7] Anne Moore has over 1 million instagram followers `https://www.instagram.com/itsannemoore/?hl=en`. Tyrone Davis was an American blues singer `https://en.wikipedia.org/wiki/Tyrone_Davis` and Allison Williams is an actress `https://en.wikipedia.org/wiki/Allison_Williams_(actress)`

Table 1: Summary statistics for Google Search Data

	Mean	Std Dev	Min	Max	Observations
Black	0.50	0.50	0	1	865
Impressions	50.9	211.4	0	3016	865
Click Through Rate	0.0055	0.025	0	0.50	741
Ad Eligible	0.17	0.37	0	1	865
Est. first page bid	11.7	4.76	0.030	29.7	864
Avg. monthly searches (000)	8.81	50.3	0.0050	550	865
Quality score reported	0.85	0.36	0	1	865

Note: Data on campaign-level.

On average, a campaign had 50.9 impressions, though there was high variation across campaigns. The average click-through across campaigns rate was 0.0055 based on campaigns that received more than one impression. Across all campaigns, the estimated first page bid was USD 11.70. By the end of our six-week period, only 16% percent of our ads were eligible to be shown, that is, were treated as live campaigns by Google and appearing next to searches. This variable is reported by the Google system to aid advertisers to understand what campaigns are being shown, and which campaigns the system has decided are not of high enough interest to users to show.

We separately downloaded from Google's keyword planner tool historic metrics on the average number of minimum and maximum monthly searches for each full name. Google does not provide us with precise data points but instead indicates a rough estimate of the search frequency (for example 10, 100 or 1000). For each name, we took the midpoint of the minimum – maximum range to give us an estimate of average monthly searches. As Table 1 indicates, across all names, average monthly searches were 8,810, though there was significant variation across names.

2.2 Descriptive Analysis of Search Advertising Data

The data that Google reports to advertisers include the variable 'status' that informs an advertiser about the likelihood of their ad being displayed in any upcoming search. Table 2 displays the values of this variable at the end of our six-week long campaign both overall and separately by campaigns targeting white-name searches and campaigns targeting Black-name searches. As part of this status update, Google reports when a 'Low Quality Score' had been assigned to a campaign. The quality score aggregates several characteristics related to an ad into a single score. A key

Table 2: Reporting what percentage of campaigns had different outcomes by Black-name and white-name searches

Status	White	Black	Total
Eligible	11.11	22.17	16.65
Low Quality Score	53.70	38.57	46.13
Low Search Volume	0.00	0.23	0.12
Not High Enough Bid	35.19	39.03	37.11
Total	100.00	100.00	100.00
Observations	865		
Observations with non-zero impressions	741		

attribute is the expected click-through rate. It also accounts for how closely an ad matches a search and characteristics related to the landing page (e.g., the bounce rate). The algorithm for the quality score is not publicly available. Google also reports per campaign whether it judged the search volume or the bid to be low.

Table 2 and Figure 2 demonstrate that at the end of this six-week period, Google judged 22% of campaigns targeted at Black-name searches but only 11% of campaigns targeted at white-name searches to eligible to be displayed when a user searched for that name (N = 865, t=4.41, $P < 0.001$). Table 3 repeats the analysis of Table 2 but excludes campaigns lacking impressions. The substantive findings are similar, but it is clear that the system labeled the zero-impression campaigns as having low volumes of searches. Crucially, though, the proportion of campaigns considered as eligible is much higher for those targeted at Black names than those targeted at white names.

We then analyzed the frequency by which ads were displayed across all campaigns. Figure 3 shows that on average a campaign targeted towards a white-name search received 88 views while a campaign targeted towards a Black-name search received 13 views. Only 16 campaigns targeted towards white-name searches but 108 campaigns targeted towards Black-name searches received zero impressions. We found that this pattern mirrors the historical average monthly search volume

Table 3: Reporting what percentage of campaigns had different outcomes by Black-name and white-name searches (excluding campaigns with zero impressions)

Status	White	Black	Total
Eligible	11.30	18.77	14.57
Low Quality Score	55.77	50.77	53.58
Not High Enough Bid	32.93	30.46	31.85
Total	100.00	100.00	100.00
Observations	865		

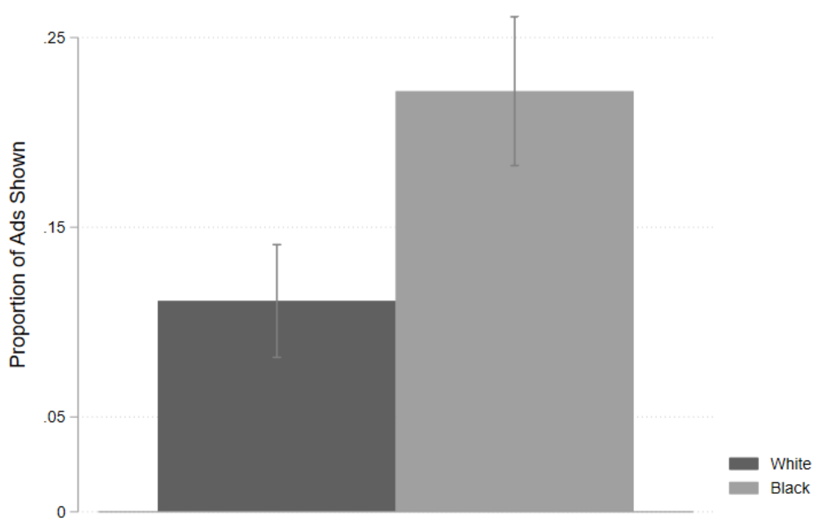

Figure 2: Ad more likely to be judged eligible to be shown alongside searches for Black names than white names (data on campaign-level)

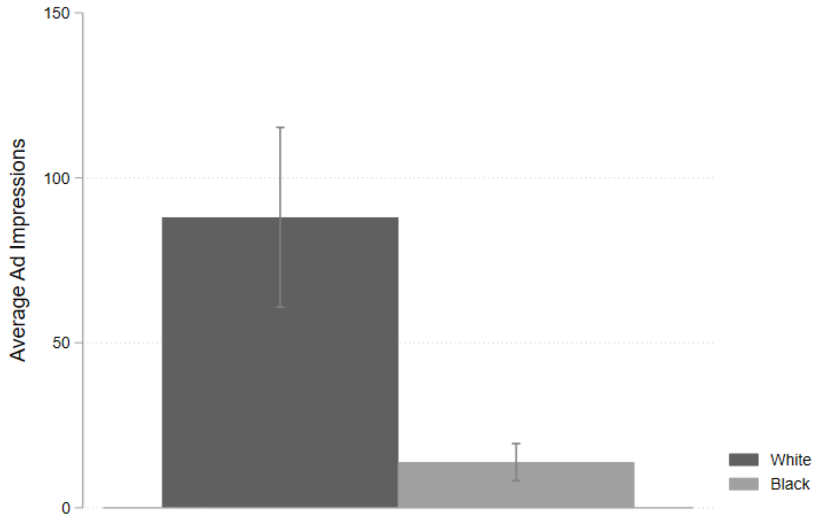

Figure 3: More ad impressions shown following white-name searches than following Black-name searches (data on campaign-level)

reported by Google for names in our data where, on average, users are more likely to search for white names than for Black names (13,010.94 vs 4618.49, $N = 865$, $t = 2.46$, $P = 0.01$). Figure 4 illustrates this difference in logarithmic terms.

To understand the lower search volume for Black names relative to white names, we turned to 1990 census data documenting the frequency of first names in the population.[8] For the names in our data, Figure 5 illustrates that the likelihood of someone having a first name typically given to white people is higher than having a first name typically given to Black people (0.08% vs. 0.02%, $N = 767$, $t = 3.29$, $P = 0.0017$). This is because there are fewer Black people than white people in the population and because, relative to white people, Black people are less likely to have common names (Fryer Jr and Levitt 2004).

By the end of the campaign, a user searching for a Black name was far more likely to be shown our ad than a user searching for a white name, but white-name searches had been significantly more frequent in the interim, likely because individual white names occur more often in the population.

[8]1990 was the last year we could find this data for. There were 7 first names where there was no frequency data, of which 6 were Black names. This suggests that these names were unusual or novel enough to have not been counted in the frequency tabulations of the 1990 census exercise.

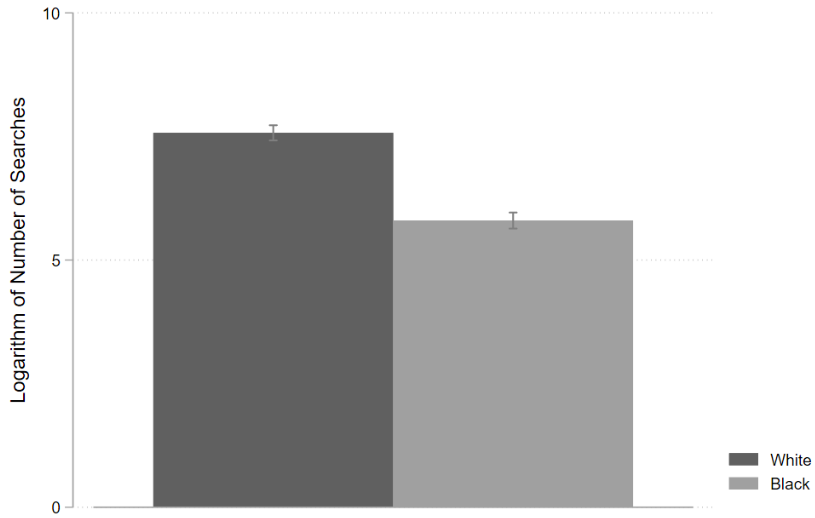

Figure 4: Historically, the number of searches for individual white names exceeds those for individual Black names (data on campaign-level)

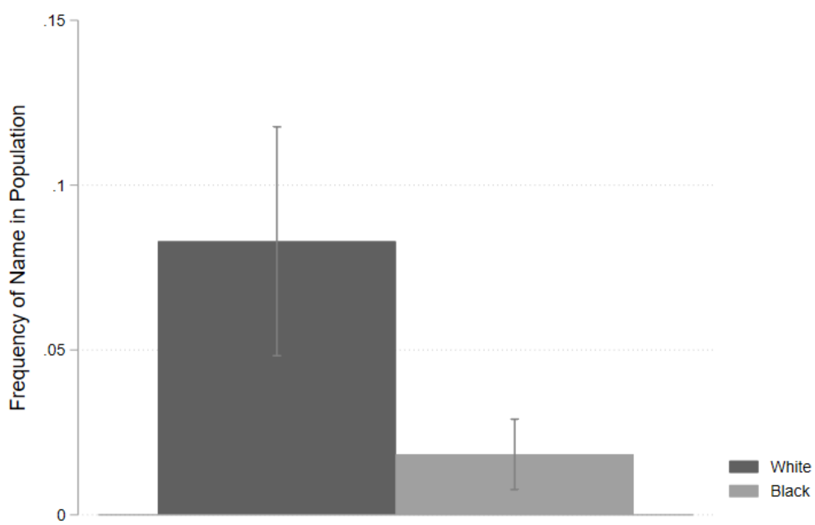

Figure 5: Lower number of impressions following Black-name searches may reflect that typically Black names are less frequent in the US population (data on name-level)

2.3 Establishing the Mechanism

We explore three possible explanations for the patterns we observe. First, we discuss whether differences in the quality score related to an ad resulting from different propensities of users to click on the same ad following Black-name and white-name searches may explain our results. Second, we turn to the algorithmic learning process that helps an algorithm to determine which ad to display. Third, we discuss potential differences in the price to display ads following Black-name and white-name searches.

2.3.1 Can Differences in Click-Through Rates Explain the Results?

It is possible that differences in the likelihood users will click on an ad might drive the results. The number of ads Google displays to a user in any individual search is limited. In order to show the most profitable and relevant ads, Google's algorithm constructs a 'quality score' that predicts the likelihood a user will click on the ad. The information on quality score and the prices an advertiser is willing to pay enter an auction mechanism that determines which ads are displayed.

The quality score plays a pivotal role in whether an ad is displayed following a targeted search. Importantly, when a campaign's ad is judged to have a low quality score relative to other ads competing in the same auction, the platform may decide that the ad is not eligible to be shown.

This quality score takes into account some factors that are common across our campaigns, and therefore cannot explain the uneven results, such as the landing page and the overall account performance. However, the quality score also relies on the propensity to click. A higher propensity to display ads at the end of the campaign following Black name searches could be a result of searchers being more likely to click when the ad followed Black name searches, a potential mechanism previously suggested by Sweeney (2013) and Barocas and Selbst (2016).

In our context, when the ad is shown, the probability of a click is generally low. The median click-through rate is 0% and the mean click-through rate is 0.55%.[9] Still, we explore possible

[9] For our discussion of average click-through rates, we always compute campaign-level click-through rates for campaigns that have more than zero impressions and then average those campaign-level click-through rates to obtain an average across campaigns. This approach is appropriate since the algorithm optimizes each campaign separately. If we compute the ratio of total number of impressions/total number of clicks by whether a search was for a white name or a Black name, this gives us an overall CTR of 0.007 for white-name searches and 0.011 for Black-name searches.

differences in click-through rates across Black and white name searches. We find that in our data, click-through rates on ads are virtually identical across campaigns (for campaigns with at least one impression: Black names 0.53% vs. white names 0.56%, $N = 741$, $t = 0.176$, $P < 0.859$). Thus, the patterns we document persist in the absence of differences in click-through rates, and the uneven outcomes do not simply reflect biases in the behavior of those who search on the platform.

2.3.2 Can Algorithmic Learning to Establish a Quality Score Explain the Results?

We explore as an alternative explanation whether the process by which an algorithm learns about the quality of an ad can result in the uneven display of ads following Black name and White name searches. An algorithm requires a minimum number of ads being shown to learn about that ad's underlying quality score. Indeed, Figure 6 illustrates that in our data, the average number of impressions in a campaign where a quality score was not reported was 0.10 relative to 60.11 impressions for campaigns where a quality score was reported ($N = 865$, $t = 3.03$, $P = 0.003$).

We therefore ask whether at the end of our campaigns, the platform had simply not yet been able to learn about users' response to ads following Black name searches, but instead had learned about user response following white name searches, leading to the distortions we observe.

We first examine the quality score the platform allocated to the campaigns. We find that as a result of the overall low click-through rates in our data, when the algorithm evaluates the quality of our ad, it is always evaluated as being low; in 54.5% of cases it has the lowest possible value of 1. This pattern demonstrates that once the algorithm had accumulated information about a campaign's ad quality, the campaign was predominantly judged as not being eligible to be shown.

We then evaluate whether there are any differences in whether quality scores are reported for campaigns targeting Black-name or white-name searches. Figure 7 demonstrates that campaigns targeted towards Black-name searches were less likely to report a quality score than campaigns targeted towards white-name searches (74% relative to 95%, $N = 865$, $t = 8.94$, $P < 0.001$).

This pattern is consistent with our earlier finding that Black names are searched for less often than white names, presumably because individual Black names are less frequent in the US population. By implication, at the end of the six-week period, the advertising algorithm was less likely to have learned about the low quality of campaigns targeting Black-name searches than about the

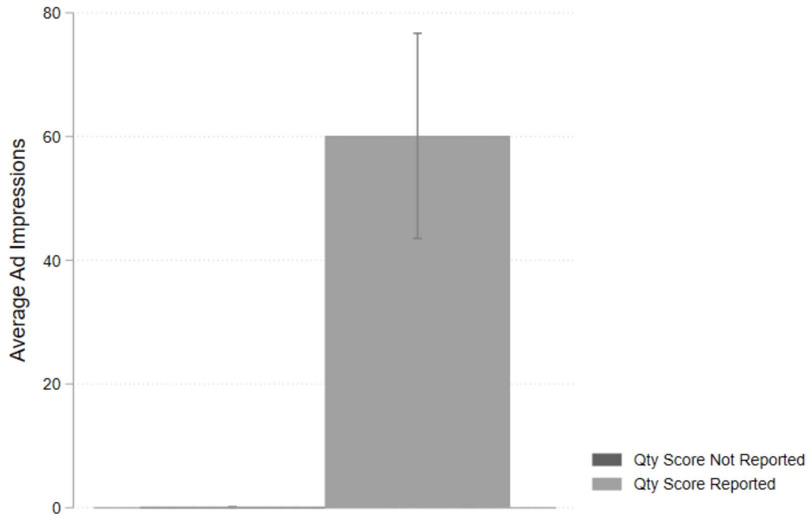

Figure 6: The more impressions an ad has, the more likely Google is to be able to record a quality score (data on campaign-level)

low quality of campaigns targeting white name searches. As a result, the algorithm was more likely to consider a campaign targeted towards users searching for Black names as being eligible to be shown.

We highlight that in our study we focus on differences in quality scores across the different keyword campaigns. Though in theory it is possible for a search engine to adjust the quality score at the account level, which would lead to a more even effect, platforms' incentives and advertisers' goals are typically such that they want to identify what works best at the most granular level.

2.3.3 Evidence for Algorithmic Learning in a Regression Analysis

We confirm these findings in regression analysis with the objective of linking our descriptive findings in Section 2.3.2 with the patterns regarding the frequency of searches and impressions established in Section 2.2.

Table 4 summarizes the results. Column (1) shows that on average, ads targeted towards Black-name searches are significantly more likely to be displayed at the end of our campaign. In Column (2), we control for the number of impressions in a campaign. It demonstrates that the effect of Black-name searches continues to hold, though the number of past impressions reduces the

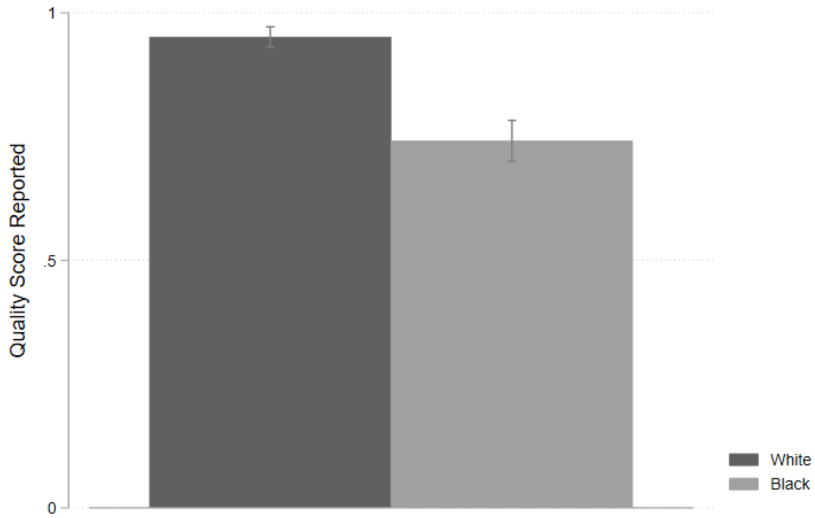

Figure 7: Ads next to Black names were far less likely to have a quality ad being reported (data on campaign-level)

likelihood of an ad to be shown. One issue with using the number of impressions as a control is that 14.3% of campaigns had zero impressions, leading that variable to have a skewed distribution.

We therefore add as an incremental control the historic measure of monthly searches. Column (3) shows that the coefficient indicating whether the campaign was targeted towards Black- or white-name searches becomes insignificant once we account for search volume, as captured by the log of the average number of historic searches. This result is consistent with the idea that, for a search term with a large number of searches, the algorithm had a greater chance to display an ad throughout the campaign, therefore accelerating algorithmic learning. Column (4) demonstrates that indeed whether the quality score was reported reduces the probability of an ad being shown.

If indeed the number of searches for a name affects algorithmic learning and, thus, whether an ad is eligible to be displayed, then whether an ad is targeted towards Black- or white-name searches should not affect its eligibility, once we hold constant the number of searches. In Column (5), we condition on the number of average historic searches being 550 (this subsample includes 220 Black and 229 white names) and in Column (6), we condition on the number of average historic searches being 5500 (subsample includes 40 Black and 156 white names). In both instances, the results confirm that eligibility does not vary with a search being for a Black or a white name.

Table 4: Eligibility for ad to be shown in a campaign

	All observations				Avg. searches 550 searches	Avg. searches 5500
	(1)	(2)	(3)	(4)	(5)	(6)
	Ad Eligible	Ad Eligible	Ad Eligible	Ad Eligible	Ad Eligible	Ad Eligible
Black	0.111***	0.103***	0.0458	0.0382	0.0325	-0.0569
	(0.0251)	(0.0254)	(0.0281)	(0.0284)	(0.0330)	(0.0521)
Impressions		-0.000105+	-0.0000587	-0.0000574		
		(0.0000602)	(0.0000604)	(0.0000604)		
Ln(Avg searches)			-0.0340***	-0.0306***		
			(0.00752)	(0.00775)		
Quality score reported				-0.0658+		
				(0.0370)		
Constant	0.111***	0.120***	0.374***	0.411***	0.127***	0.107***
	(0.0177)	(0.0185)	(0.0589)	(0.0624)	(0.0231)	(0.0234)
Observations	865	865	865	865	449	199
R-Squared	0.0220	0.0255	0.0482	0.0516	0.00215	0.00602

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Average monthly searches Google in thousands. Data on campaign-level.

Consequently, our data provide evidence that the mechanics inherent in algorithmic learning, a process which is required for an algorithm to then make optimal decisions, can lead to uneven outcomes in the types of ads being displayed in response to searches related to members of different racial groups. While our study uses an innocuous ad relating to government jobs, these algorithmic learning patterns may lead to disparate impacts in protected sectors, as demonstrated by Sweeney (2013).

Our analysis in Table 4 includes campaigns that received zero impressions during our observation period. This is because we are trying to understand the likelihood of someone seeing a campaign if they were to search for that name, rather than the performance of any campaign. An ad impression count of zero still means that in theory the ad will be displayed if someone sees that ad. Web Appendix Table A3 excludes observations with zero impressions and show similar patterns. This is comforting as it ensures we avoid a purely mechanical result due to campaigns that necessarily remain eligible.

2.3.4 Can Differences in Advertisers' Willingness-to-Pay Explain the Results?

Last, we turn to the possibility that differences in the willingness-to-pay of different advertisers may cause the patterns we observe. The maximum bid an advertiser specifies relative to the maximum bid specified by competitors plays an important role in determining whether an ad is shown. Indeed, Lambrecht and Tucker (2019) show that differences in competitors' willingness-to-pay can lead to ads for information on STEM careers being less likely to be displayed to women than to men.

As indicated in Table 2, for 37.1% of campaigns, the platform reports our bid as not being high enough. Importantly, however, this rate does not differ significantly between Black- and white-name searches (39.0% vs 35.2%, $N = 865$, $t = 1.17$, $P = 0.242$). Still, we explore whether the price other advertisers were willing to pay for displaying an ad affects our results.

For this purpose, we collected data on the estimated first page bid reported by Google AdWords.[10] This variable measures how much other advertisers value a targeted keyword and, therefore, allows us to measure whether what we observe is primarily a pricing effect. Note that it is not straightforward how such a mechanism would explain the complex pattern we observe: If other advertisers bid higher when targeting white-name searches, our uniform bid could possibly lead to the platform being less likely in the future to display our ads following white-name searches. However, such a mechanism could not explain the high number of impressions for white-name searches we documented in Section 2.2. Conversely, if other advertises were willing to pay less in campaigns targeting white-name searches, this could explain why, throughout the six weeks that our campaigns ran, our ad with a uniform bid was displayed significantly more frequently following white-name searches. However, that pattern would not rationalize why our campaign ads were less likely to be shown following such white-name searches at the end of the data period.

Nonetheless, we explore differences in the reported first page bids.[11] We find that the estimated first page bid is higher for white-name than for Black-name searches (12.15 vs. 11.21, $N = 864$, $t = 2.90$, $P = 0.004$). We then, in Column (1) of Table 5, control for estimated first page bid in addition to the indicator 'Black', not controlling for variables related to search volume. Unsurprisingly, the probability that our ad was displayed, given our maximum bid, declines in the first page bid estimate. This control adds significant explanatory power to the estimation, because very high competitive bids make it extremely unlikely for our ad to be displayed. However, while the coefficient for 'Black' is somewhat lower, it is still sizable and highly significant, suggesting that

[10]Such data on estimated first page bids has previously been used to understand price patterns in online search (Goldfarb and Tucker 2011).

[11]Google did not provide an estimate for the search term Hakim Miller.

Table 5: Eligibility for ad to be shown in a campaign – Accounting for bids

	(1) Ad Eligible	(2) Ad Eligible	(3) Ad Eligible
Black	0.0651**	0.0276	0.0163
	(0.0214)	(0.0238)	(0.0240)
Impressions	-0.0000859+	-0.0000609	-0.0000585
	(0.0000505)	(0.0000522)	(0.0000519)
Ln(avg. searches)		-0.0222***	-0.0171*
		(0.00649)	(0.00667)
Quality score reported			-0.0958**
			(0.0312)
Est. first page bid	-0.0424***	-0.0418***	-0.0421***
	(0.00222)	(0.00224)	(0.00223)
Constant	0.634***	0.792***	0.849***
	(0.0310)	(0.0546)	(0.0574)
Observations	864	856	856
R-Squared	0.317	0.326	0.334

+ p <0.10, * p <0.05, ** p <0.01, *** p <0.001. Average monthly searches Google in thousands. Data on campaign-level.

the addition of the estimated first page bid explains differences across Black-name and white-name searches to only a limited extent.

In Columns (2) and (3) we then control for the full set of variables previously included, that is, those relating to the number of searches and whether a quality score was reported. As expected and consistent with Columns (3) and (4) in Table 4 the indicator for 'Black' now becomes insignificant. Again, Web Appendix Table A4 excludes observations with zero impressions and shows similar patterns.

In sum, our results demonstrate that advertisers' willingness-to-pay plays only a small part in explaining the racial differences we documented. This suggests that, unlike in other work such as Lambrecht and Tucker (2019), ad pricing is not the main factor driving our result.

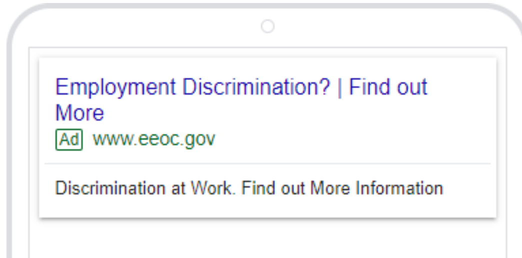

Figure 8: Ad creative

2.3.5 Does the Number of Competitors Affect Results?

We obtained from Google data on competitors that were advertising for the same keywords at the same time as we did. We classify competitors as public record companies or as other competitors. In Table A5 in our appendix we demonstrate that neither including in our regression the number of competitors that are public record companies nor including the number of other competitors that advertise at the same time as we do shifts the results. Section A.3 in the Online Appendix discusses this analysis.

3 Extending the Result to the Context of Religion

A natural question is whether this mechanism extends to other contexts. Therefore we run a similar experiment to establish how the phenomenon applies to related contexts. Specifically, we focus on online searching for information related to religious discrimination. We implemented search advertising campaigns on Google AdWords. We instructed Google AdWords to target an ad to users searching the keywords "discrimination 'religion'" where 'religion' was a placeholder for the eleven most common religious denominations in the US: Atheist, Buddhist, Catholic, Evangelical, Hindu, Jehovah's Witness, Jewish, Mormon, Muslim, Orthodox and Protestant. We additionally included the term 'Christian' to refer more generally to Christian faiths. For example, our ad would be shown when someone used the search term 'discrimination jewish.' Such search terms may be targeted by lawyers seeking clients for lawsuits, for example in the context of employment discrimination. The ad offered information on employment discrimination and was identical throughout campaigns. Figure 8 displays the creative. The ad linked to a government website that provided practical information about employment law.

We instructed Google AdWords to use 'broad match,' which means that Google considered the ad in a search auction when the specific term was used in the search query as well as when similar terms were used. For example, our ad would have been shown when someone used precisely the search term 'discrimination muslim' and also when someone searched for 'some ways how muslim people are discriminated at work.' Two reasons motivate our choice of broad match. First, unlike what might be the case for name search, slight deviations from the precise search terms do not typically imply a different topical interest. Second, not requiring the exact wording means that we can target a larger number of searches and thus collect data more quickly. We set a maximum daily campaign budget of $100.

Table 6 shows that after one day of running the ads, Google AdWords gave a low quality score for the campaigns targeting 'discrimination jewish' and 'discrimination muslim' after having shown 18 and 20 impressions. As a result, these campaigns were no longer marked as 'eligible' to be shown. Campaigns targeting the remaining faiths (Atheist, Buddhist, Catholic, Evangelical, Hindu, Jehovah's Witness, Mormon, Orthodox, Protestant) each received between 0 and 2 impressions and continued to be eligible for showing our ad. The campaign targeting the broader term 'discrimination Christian' had received 25 impressions and continued to be eligible for showing the ad. This campaign differed because it did not specify a particular religious group but referred to a broader affiliation and had a higher click-through rate (8.0% relative to 5.56% and 0%). This pattern suggests that any algorithmic learning process can potentially start at relatively low numbers of impressions.

This study provides two insights. First, it demonstrates in a different empirical context that the process of algorithmic learning in online advertising can affect the minority and the majority group in different ways. In this setting, the majority and minority group reflects the amount of searching that is being done. So even though there are more protestants in the US population than muslims, perhaps because of historic privilege fewer protestants experience employment discrimination. Again this shows, that what matters is digital participation by groups, rather than baseline population levels. Second, while study 1 documented differences in the probability of an ad being shown after a period of six weeks, study 2 demonstrates that the process of algorithmic learning

Table 6: Overview of results, study in the context of religious discrimination

Keyword: Discrimination + ...	Status	Clicks	Impr.	Impr. share (%)
Atheist	Eligible	0	1	<10
Buddhist	Eligible	0	0	<10
Catholic	Eligible	0	2	<10
Evangelical	Eligible	0	0	–
Hindu	Eligible	0	0	<10
Jehovah's Witness	Low Search Volume	0	0	–
Jewish	Low Quality Score	1	18	13.64
Mormon	Eligible	0	0	<10
Muslim	Low Quality Score	0	20	15.04
Orthodox	Eligible	0	0	–
Protestant	Eligible	0	0	–
Christian	Eligible	2	25	<10

can start to produce differences in outcomes even after a short time period.

4 Summary, Discussion and Limitations

4.1 Summary

In this research, we ask empirically whether the simple mechanics by which real-time algorithms operate can lead to outcomes that disadvantage minority groups. To explore empirically why and whether such patterns occur, we carried out two field tests from the perspective of an advertiser. Our first field test implemented an advertising campaign targeting a large number of names that are typically used by either Black or white people. We show that for advertising campaigns targeting Black-name searches, an algorithm accumulates information more slowly than for campaigns targeting white-name searches simply because Black names are searched for less frequently, presumably because they are less common in the population. As a result, the algorithm learns about the quality of the underlying ad more slowly and an ad is more likely to persist for searches next to Black names. This holds despite people being no more likely to click on the ad accompanying a Black-name search than if the same ad accompanies a white-name search. Evaluating algorithmic fairness through the percentage of individuals affected is consistent with a legal literature (e.g., Hellman (2020)). Our results show the need to focus on whether people at a certain point in time are treated equally, rather than focusing only on whether over a period in aggregate, outcomes may be fair.

We believe that our results are important for two reasons. First, if ads shown in response to searches for minority groups are more likely to show disadvantageous content, there is the risk that such ads may on a broader societal level unintentionally reinforce negative stereotypes. Second, a slower pace of an algorithm in responding to changes over time may lead to access to new opportunities not being equally distributed. Overall, we empirically demonstrate that the seemingly innocuous process of algorithmic learning can inadvertently disadvantage minority groups. As far as we are aware, this research is the first to demonstrate the role that algorithmic learning plays in online advertising and the unintended consequences of that role.

4.2 Implications

Our findings have practical implications for advertisers. The first is simply to encourage awareness of the distortions that uniform requirements for data imposed by a platform to resolve cold-start problems can create for campaigns. This means that in a setting like paid search, where advertisers can explicitly set up what looks like a balanced campaign, these ads may not be shown equally. Advertisers should carefully monitor throughout a campaign whether in effect the algorithm is showing ads equally, even after the initial setup of intentionally balanced campaigns. This is particularly the case if the campaign is targeting variables which may be highly correlated with protected characteristics, such as, in our case, names of individuals. For example, if the number of people who reside in geographic regions varies by race, and that geographic region is used as a targeting variable, then this could lead to uneven outcomes. We first recommend that advertisers think about whether their targeting segments are likely to be exposed to similarly sized populations. If they do, then there are unlikely to be issues from algorithmic learning. However, if they are uneven, advertisers should use the granular data available to them from advertising dashboards that platforms provide to advertisers to check whether algorithmic learning is leading to distortions. This is not just the case if the advertiser is using a variable that is potentially correlated with a sensitive variable for targeting, but also if the advertiser is selling products that themselves are sensitive due to their welfare implications - such as health, education or financial products. In each case, differential speeds of algorithmic learning might affect the quality of recommendations available.

Our findings have implications for digital platforms. While the platform's goal in using algorithmic learning may be to ensure that consumers only see ads they are interested in, our results suggest the platform needs to be aware of the possible uneven effects resulting from such tools. While a platform cannot intervene in an individual advertiser's campaign, it may want to educate advertisers about challenges related to uneven outcomes, such as that different rates of exposure across similar campaigns may lead to disparate treatment of different social groups. It also suggests that the use of algorithmic learning to try to address the cold-start problem inherent in content environments where quality is uncertain (Claussen et al. 2024), may itself be problematic. In particularly sensitive environments such as those related to protected characteristics, the cold-start problem may need to be reanalyzed to see if there are other ways of addressing it, such as pooling data across customers. In particular, platforms should consider whether setting a standardized threshold for data collection to resolve the cold-start problem is always desirable, especially in circumstances where either the product or the nature of targeting itself is sensitive. Alternatively, platforms can consider the extent to which algorithms should leverage insights across different groups targeting by the same or very similar content.

Our findings have implications for public policy. Governments throughout the world have wrestled with how to address the possibility that algorithms might reinforce inequality. Several policy approaches have been suggested, including algorithmic transparency and algorithmic auditing.[12] However, such policies tend to presuppose a static process of algorithmic determination where an algorithm makes predictions on the basis of an established set of training data. Our findings differ from the more typical concern that such a training data set exhibits 'sample size disparity' in three ways. First, in our context, the unevenness arises from the speed by which new data are fed into a real-time learning process instead of from differences in a static data set. Second, individual data points are contributed by a large number of independent agents over time, and therefore it is not possible for a single agent to 'correct' the unevenness in data *ex ante* in order to generate more even outcomes. Third, prior research worried about settings where there was unrepresentative or

[12] https://www.accc.gov.au/system/files/Digital%20platforms%20inquiry%20-%20final%20report.pdf

insufficient data for each group to allow the algorithm to make even decisions once it had been trained. By contrast, in our setting the uneven outcomes arise during the learning phase, *because* the data is representative.

Given that algorithmic learning is a ubiquitous tool used in real-time environments, it is difficult to restrict such a process. A more practical way of addressing this challenge may be identifying specific empirical advertising contexts where algorithmic learning may be particularly harmful to social groups, and advise digital platforms to pool data across consumers in a way which can mitigate the uneven display of digital content. Though minority groups being associated with different digital content may not seem directly harmful, it is important to remember that racial disparity is often "the product of countless, mostly unconscious daily procedures and decisions."[13] In our minds, this paper stresses that seemingly innocuous processes can affect minority groups in significant ways – which, even beyond any detrimental impact for an individual, can create a broader environment and society that appears hostile to minority groups.

4.3 Limitations

There are, of course, limitations to our research. First, our main study is in part motivated by the finding of Sweeney (2013) that undesirable ads are more likely to be shown following the search for a Black relative to a white name. We provide evidence that uneven speeds in algorithmic learning contribute to such outcomes. However, it is still possible that other factors contribute to the patterns reported by Sweeney (2013), such as advertisers' deliberate policies. Beyond documenting the persistent pattern of background check advertising practices, we do not have insight into internal policies. Though we control for obvious differences such as the number of competitors and clicks, we do not control for everything, such as ad prices faced by these competitors. Second, the precise implication of the effect we document for inequality will depend on whether the content is positive or negative, and whether the smaller group itself is advantaged or disadvantaged. We emphasize that while our study is focused on a setting where algorithmic learning negatively affects minority groups, either because the display of harmful content may hurt them or because they are less likely to be exposed to beneficial information. However, we acknowledge that it is likewise possible that the

[13] `https://www.ft.com/content/baf58652-c511-4556-8ae3-0cd79c06117a`

process of algorithmic learning may at times benefit minority groups. We leave the exploration of this topic to future research. Third, while we document that algorithmic learning can inadvertently disadvantage minority groups, our paper does not attempt to suggest specific algorithmic designs that would circumvent this problem. Notwithstanding these limitations, we believe our paper is a useful first step in documenting the role of algorithmic learning in causing differential effects among minority groups.

References

Abu-Elyounes, D. (2020). Contextual fairness: A legal and policy analysis of algorithmic fairness. *University of Illinois Journal of Law, Technology, & Policy*, 1.

Acquisti, A. and C. Fong (2020). An experiment in hiring discrimination via online social networks. *Management Science 66*(3), 1005–1024.

Ali, M., P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, and A. Rieke (2019, nov). Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes. *Proc. ACM Hum.-Comput. Interact. 3*(CSCW).

Ali, M., P. Sapiezynski, A. Korolova, A. Mislove, and A. Rieke (2021). Ad delivery algorithms: The hidden arbiters of political messaging. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, New York, NY, USA, pp. 13–21. Association for Computing Machinery.

Barocas, S., M. Hardt, and A. Narayanan (2017). Fairness in machine learning. *NIPS Tutorial 1.*

Barocas, S. and A. D. Selbst (2016). Big data's disparate impact. *Californian Law Review 104*, 671.

Bertrand, M. and S. Mullainathan (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review 94*(4), 991–1013.

Burroughs, V. J., R. W. Maxey, and R. A. Levy (2002). Racial and ethnic differences in response to medicines: towards individualized pharmaceutical treatment. *Journal of the National Medical Association 94*(10 Suppl), 1.

Claussen, J., C. Peukert, and A. Sen (2024). The editor and the algorithm: Returns to data and externalities in online news. *Management Science 0*(0), 0.

Cowgill, B. (2017). Automating judgement and decision-making: Theory and evidence from résumé screening. In *Columbia University, 2015 Empirical Management Conference.*

Cowgill, B. (2018). The impact of algorithms on judicial discretion: Evidence from regression discontinuities. Technical report, Working paper.

Dastin, J. (2018). Amazon scraps secret ai recruiting tool that showed bias against women. *San Fransico, CA: Reuters. Retrieved on October 9*, 2018.

Datta, A., M. C. Tschantz, and A. Datta (2015). Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies 2015*(1), 92–112.

Dressel, J. and H. Farid (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances 4*(1).

Edelman, B., M. Ostrovsky, and M. Schwarz (2007). Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American Economic Review 97*(1), 242–259.

Fryer Jr, R. G. and S. D. Levitt (2004). The causes and consequences of distinctively black names. *The Quarterly Journal of Economics 119*(3), 767–805.

Ghose, A. and S. Yang (2009). An empirical analysis of search engine advertising: Sponsored search in electronic markets. *Management Science 55*(10), 1605–1622.

Goldfarb, A. and C. Tucker (2011). Search engine advertising: Channel substitution when pricing ads to context. *Management Science 57*(3), 458–470.

Hellman, D. (2020). Measuring algorithmic fairness. *Va. L. Rev. 106*, 811.

Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer (2015). Prediction policy problems. *American Economic Review 105*(5), 491–95.

Lambrecht, A. and C. Tucker (2019). Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management science 65*(7), 2966–2981.

Mitchell, S., E. Potash, S. Barocas, A. D'Amour, and K. Lum (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application 8*.

Nachbar, T. B. (2020). Algorithmic fairness, algorithmic discrimination. *Florida State University Law Review 48*, 509.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366*(6464), 447–453.

Rutz, O. J. and R. E. Bucklin (2011). From Generic to Branded: A Model of Spillover Dynamics in Paid Search Advertising. *Journal of Marketing Research 48*(1), 87–102.

Srinivasan, R. and G. Sarial-Abi (2021). When algorithms fail: Consumers' responses to brand harm crises caused by algorithm errors. *Journal of Marketing 85*(5).

Sweeney, L. (2013). Discrimination in online ad delivery. *ACMQueue 11*(3), 10.

Tunuguntla, S. and P. R. Hoban (2021). A near-optimal bidding strategy for real-time display advertising auctions. *Journal of Marketing Research 58*(1), 1–21.

Ukanwa, K. and R. T. Rust (2021). Algorithmic discrimination in service.

5 Appendix

A.1 Recap of Results by Sweeney (2013)

Figure A1 shows the percentage of Black- and white-name searches in response to which public record ads were displayed in Sweeney (2013)'s original research (based on Figure 16 in the paper). Though Sweeney (2013) also discusses the distribution of ads on Reuters, we focus in this research on Google search ads so this figure reports the results for Google only. It is clear that the probability of a public record ad being displayed was higher for Black-name searches than for white-name searches.

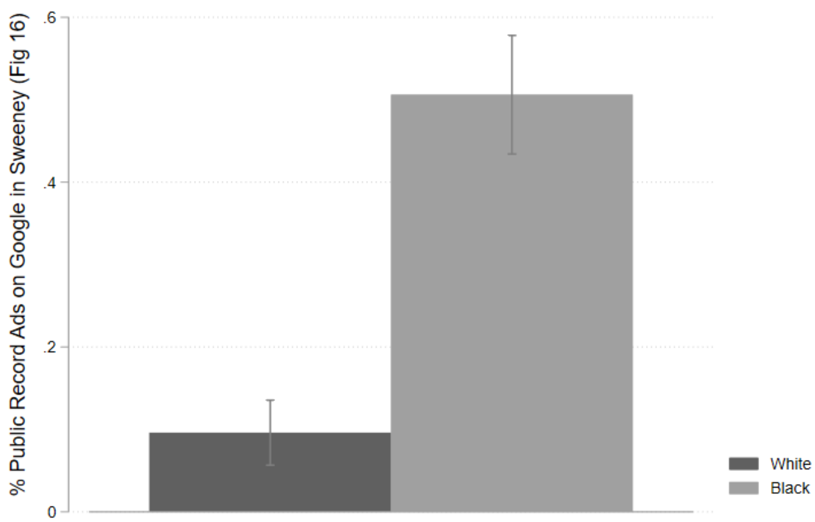

Figure A1: Percent of public record ads displayed in response to Black-name and White-name searches

A.2 Additional Tables

Here, we report additional Appendix Tables that the main paper refers to, including tables on the names used and robustness checks of the empirical results when excluding observations with zero impressions.

A.3 Insights from Competitive Intelligence

One motivation of our study was Sweeney (2013) who had demonstrated that ads for background checking services were more likely to be shown following searches for Black than for white names. In our study, we purposely did not show an ad for background checking services, but uncovered similar patterns for a different type of ad. However, we can use the data that Google reports on competitive bidders to shed light on the extent to which background checking services bid for ads towards Black or white names when we advertised.

Table A1: First names used

Black Female	*White Female*	*Black Male*	*White Male*
Aaliyah	Allison	Darnell	Brad
Aisha	Amy	Deandre	Brendan
Deja	Anne	Deshawn	Brett
Ebony	Carrie	Hakim	Cody
Imani	Claire	Jamal	Connor
Keisha	Emily	Jermaine	Dustin
Kenya	Emma	Kareem	Geoffrey
Lakisha	Jill	Leroy	Greg
Latanya	Katelyn	Malik	Jack
Latisha	Katie	Marquis	Jake
Latonya	Kristen	Rasheed	Jay
Latoya	Laurie	Terrell	Luke
Nia	Madeline	Tremayne	Matthew
Precious	Meredith	Trevon	Neil
Shanice	Molly	Tyrone	Tanner
Tamika			Wyatt

Table A2: Top 20 last names from 2010 Census

	Percent White	Percent Black	> 90% Hispanic
Anderson	75.2	18.9	0
Brown	58	35.6	0
Davis	62.2	31.6	0
Garcia	5.4	.5	1
Gonzalez	4	.4	1
Hernandez	3.8	.4	1
Jackson	39.9	53	0
Johnson	59	34.6	0
Jones	55.2	38.5	0
Lopez	4.9	.6	1
Martin	74.8	15.8	0
Martinez	5.3	.5	1
Miller	84.1	10.8	0
Moore	66.4	27.7	0
Rodriguez	4.8	.5	1
Smith	70.9	23.1	0
Taylor	65.4	28.4	0
Thomas	52.6	38.8	0
Williams	45.8	47.7	0
Wilson	67.4	26	0
Total	45.255	21.67	.3

Table A3: Eligibility for ad to be shown in a campaign, excluding observations with zero impressions

	All observations				Avg. searches 550 searches	Avg. searches 5500
	(1)	(2)	(3)	(4)	(5)	(6)
	Ad Eligible	Ad Eligible	Ad Eligible	Ad Eligible	Ad Eligible	Ad Eligible
Black	0.0747**	0.0673*	0.0339	0.0339	0.0384	-0.0539
	(0.0260)	(0.0263)	(0.0283)	(0.0283)	(0.0355)	(0.0516)
Impressions		-0.000102+	-0.0000710	-0.0000711		
		(0.0000575)	(0.0000580)	(0.0000581)		
Ln(Avg searches)			-0.0242**	-0.0242**		
			(0.00788)	(0.00789)		
Quality score reported				0.00867		
				(0.102)		
Constant	0.113***	0.122***	0.304***	0.295*	0.132***	0.104***
	(0.0172)	(0.0180)	(0.0616)	(0.116)	(0.0241)	(0.0234)
Observations	741	741	741	741	408	194
R-Squared	0.0110	0.0153	0.0277	0.0277	0.00288	0.00565

+ p <0.10, * p <0.05, ** p <0.01, *** p <0.001. Average monthly searches Google in thousands. Data on campaign-level.

Table A4: Eligibility for ad to be shown in a campaign, accounting for bids, excluding observations with zero impressions

	(1)	(2)	(3)
	Ad Eligible	Ad Eligible	Ad Eligible
Black	0.0405+	0.0170	0.0169
	(0.0235)	(0.0255)	(0.0255)
Impressions	-0.0000823	-0.0000662	-0.0000659
	(0.0000513)	(0.0000532)	(0.0000533)
Est. first page bid	-0.0376***	-0.0372***	-0.0372***
	(0.00272)	(0.00276)	(0.00276)
Ln(avg. searches)		-0.0160*	-0.0159*
		(0.00720)	(0.00721)
Quality score reported			-0.0240
			(0.0912)
Constant	0.574***	0.690***	0.713***
	(0.0364)	(0.0621)	(0.108)
Observations	741	733	733
R-Squared	0.218	0.223	0.223

+ p <0.10, * p <0.05, ** p <0.01, *** p <0.001. Average monthly searches Google in thousands. Data on campaign-level.

Keyword: [Latoya Williams] ADD FILTER

SEGMENT COLUMNS DOWNLOAD EXPANI

Display URL domain	Impression share	Avg. position	Overlap rate	↓ Position above rate	Top of page rate	Abs. Top of page rate	Outranking share
beenverified.com	19.36%	2.2	36.68%	5.48%	47.37%	15.04%	28.38%
mylife.com	10.92%	2.8	21.11%	4.76%	44.00%	14.67%	28.68%
truthfinder.com	15.43%	2.8	25.63%	3.92%	17.92%	1.89%	28.68%
spokeo.com	< 10%	3.2	13.57%	3.70%	50.00%	15.38%	28.82%
peoplefinders.com	13.25%	3.6	25.63%	1.96%	12.09%	0.00%	28.82%
peoplelooker.com	11.64%	3.5	20.60%	0.00%	30.00%	2.50%	28.97%
You	28.97%	1.1	–	–	77.89%	76.88%	–

Figure A2: Information on competitive bidders reported by Google AdWords

Google reports to advertisers how often specific competitors' ads were shown alongside their ad. Figure A2 shows a screenshot as an example. We collected such information on other advertisers who were bidding on that keyword for each of our campaigns. This set of analyses focuses on campaigns where the number of impressions was large enough for Google to report what they refer to as an 'auction insight report.' As a result, 113 campaigns targeting Black-name searches and 27 campaigns targeting white-name searches that had low search activity during the campaigns are excluded from our analysis.

First, we study the extent to which public record companies compete with our campaign.[14] We find on average across campaigns, 2.5 such competitors for Black-name searches and 2.0 for white-name searches ($N = 726$, $t = 3.52$, $P < 0.001$). This difference in the number of competing public record companies is reflected in the overall number of competitors recorded for a name. White-name searches have on average 3.3 and Black-name searches 3.8 competitors ($N = 546$, $t = 3.14$, $P < 0.002$). The number of competing advertisers that are not public record companies is not significantly different (0.610 for white-name and 0.561 for Black-name searches, $N = 546$, $t = 0.72$, $P = 0.473$).

Second, we study the share of impressions that across campaigns goes to each of the public record companies that advertise. Again, we find that for Black-name searches, any of the public record companies that advertised had, on average, 18.3% of impressions, while for white-name searches these were 10.3% ($N = 1630$, $t = 12.22$, $P < 0.001$). Google does not provide precise information on impression shares less than 10%. Hence, we set the value for impression shares between 0 and 10% to 0.05. When alternatively using values of 0.01, of 0.09, or excluding those observations from the analysis, we similarly obtain that the share of impressions a public record company has when targeting Black-name searches is significantly higher than for white-name searches.

[14]We focus on records where there was data on at least one competitor available of the type that the respective test analyzes.

Table A5: Including Presence of Competitors in Our Specification

	(1) Ad Eligible	(2) Ad Eligible	(3) Ad Eligible	(4) Ad Eligible	(5) Ad Eligible	(6) Ad Eligible	(7) Ad Eligible
Black	0.111***	0.0714***	0.0740**	0.0705**	0.0636*	0.0286	0.0285
	(0.0251)	(0.0211)	(0.0265)	(0.0267)	(0.0270)	(0.0290)	(0.0290)
Est. first page bid		-0.0425***					
		(0.00222)					
Public Record Competitors				0.00737	0.00624	-0.000129	-0.0000879
				(0.00708)	(0.00710)	(0.00732)	(0.00733)
Non Public Record Competitors				0.0157	0.0164	0.0187	0.0188
				(0.0189)	(0.0189)	(0.0189)	(0.0189)
Impressions					-0.000101+	-0.0000739	-0.0000736
					(0.0000583)	(0.0000602)	(0.0000603)
Ln(avg. searches)						-0.0261**	-0.0260**
						(0.00856)	(0.00857)
Quality score reported							-0.0240
							(0.112)
Constant	0.111***	0.628***	0.116***	0.0941***	0.105***	0.313***	0.336**
	(0.0177)	(0.0308)	(0.0176)	(0.0226)	(0.0235)	(0.0708)	(0.129)
Observations	865	864	726	726	726	719	719
R-Squared	0.0220	0.314	0.0107	0.0142	0.0182	0.0294	0.0295

+ p <0.10, * p <0.05, ** p <0.01, *** p <0.001.

Third, Google reports how much our campaigns overlapped with ads by competitors. We find that the average overlap rate with public record companies for Black-name searches was 27.2% and for white-name searches was 21.0% ($N = 1630$, $t = 4.96$, $P < 0.001$).

These results suggest that we observe a similar pattern of focus by background record companies, in that their ads are more likely to appear next to Black names, as documented by (Sweeney 2013). We also checked the robustness of our results to the presence of competitors but our results did not qualitatively change. The results of this specification are reported as Table A5.