Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)

Auditing Race and Gender Discrimination in Online Housing Markets

Joshua Asplund,[1] Motahare Eslami,[1] Hari Sundaram,[1] Christian Sandvig,[2] Karrie Karahalios[1]
[1]University of Illinois Urbana-Champaign, [2]University of Michigan
{asplund3, eslamim2, hs1, kkarahal}@illinois.edu, csandvig@umich.edu

Abstract

While researchers have developed rigorous practices for offline housing audits to enforce the US Fair Housing Act, the online world lacks similar practices. In this work we lay out principles for developing and performing online fairness audits. We demonstrate a controlled sock-puppet audit technique for building online profiles associated with a specific demographic profile or intersection of profiles, and describe the requirements to train and verify profiles of other demographics. We also present two audits using these sock-puppet profiles. The first audit explores the number and content of housing-related ads served to a user. The second compares the ordering of personalized recommendations on major housing and real-estate sites. We examine whether the results of each of these audits exhibit indirect discrimination: whether there is correlation between the content served and users' protected features, even if the system does not know or use these features explicitly. Our results show differential treatment in the number and type of housing ads served based on the user's race, as well as bias in property recommendations based on the user's gender. We believe this framework provides a compelling foundation for further exploration of housing fairness online.

Introduction

Online tracking and algorithmic personalization have opened up complicated new legal, ethical, and policy questions regarding bias and discrimination. As more and more information from users becomes available online, many platforms employ personalization algorithms to target information to users based on their behavior. While this might improve the efficiency of information consumption by users, it can also introduce bias along many dimensions including discrimination against a specific group of users (e.g. based on race and gender) (Datta et al. 2017). The effects of such biases can perpetuate inequality, and even constitute illegal conduct in some areas (Austin Turner et al. 2013).

Housing is one area where discrimination based on a protected class (including gender, race, color, disability, religion, familial status, or national origin) is subject to fairness regulations. In the United States of America, the Fair Housing Act (42 U.S.C. §3601-3619), often abbreviated FHA, makes it illegal for a landlord or seller to refuse to sell, rent to, or negotiate with a person due to their inclusion in a protected class.

In order to enforce these housing laws, the U.S. Department of Housing and Urban Development (HUD) performs regular audits to determine whether potential buyers or renters received differential treatment based on one of these protected classes. These studies are generally structured as a paired test (Wienk 1979; Schwartz 2014). In this style of audit, two testers pose as prospective buyer or renter who differ only in protected classes. The participants reach out to real-estate agents and landlords asking to be shown houses on the market that fit specific criteria, and record which properties they are shown and how they are treated.

The internet, however, has changed the way people search for housing, making traditional audits inadequate. Users now have access to listing information that was previously only available to real-estate agents or landlords. Significant portions of the housing search can now take place through property listing sites, housing search engines, and targeted advertising. Each of these can introduce subtle bias. For example, *Ragin v. New York Times* (923 F.2d 995 (1991)), ruled that housing ads depicting African American solely in service positions violated the FHA. Similarly, advertising housing only to a racially homogeneous neighborhood would be unfair.

In this work we take steps towards building online housing audits to investigate gender-based and race-based discrimination in the context of online housing advertisements and search-result ranking. We demonstrate a sock-puppet browsing technique for building agents associated with a specific demographic profile, and discuss methods of validating that these profiles are classified correctly by advertising networks. We train profiles using this system to perform two audits of online housing systems. In the first audit we examine whether there is correlation between number of housing ads served to users and their protected features, regardless of the system's knowledge or use of these features. Our results show significant differences in the number of housing advertisements shown to agents of different races, with Caucasian users receiving more ads than non-Caucasian. In the second audit, we analyze the order of listing recommendations, and show that a major property listing site orders recommended listings differently depending on a visitor's gender, presenting more expensive listings higher up for women. We also show unequal recommendations that could lead to geographical discrimination on the basis of race. Finally, we discuss how the framework outlined in this paper can provide a basis for further exploration and audits.

Related Work

This work focuses on audits, defined here as systematic evaluations of organizations, networks, or marketplaces. Audits are generally performed by independent researchers, treating the test subject (e.g. organization, network, or marketplace) as a black box. These evaluations aim to either prove that the system being tested meets formal standards or (in social science) to measure discrimination in general. In this paper we are concerned with whether online housing markets provide fair outcomes for users regardless of gender or race.

Conceptualizing Fairness

The first step in this process is conceptualizing fairness and bias. A recent survey (Narayanan 2018) identified more than 20 definitions of fairness. Many of these fit into larger categorizations by the method, target, or purpose of the discrimination. Previous auditing studies have particularly looked at two categorizations of fairness: *direct* vs. *indirect* and *individual* vs. *group* fairness (Datta, Tschantz, and Datta 2014; Chen et al. 2018; Hardt, Price, and Srebro 2016; Zemel et al. 2013).

In the *direct/indirect* categorization of fairness, the focus is on the method of discrimination (Pedreschi, Ruggieri, and Turini 2008). Direct discrimination occurs when rules or procedures explicitly impose "disproportionate burdens" on minority or disadvantaged groups. Indirect (or systematic) discrimination is caused by rules or procedures that impose disproportionate burdens without explicitly targeting the specific groups (Collins and Khaitan 2018). In algorithmic systems, this categorization of fairness concerns whether the output of the system correlates with some protected attributes, even if these attributes are not used as inputs to the system.

Individual/group fairness is based on how subjects are selected from the groups being studied (Dwork et al. 2012). Individual fairness requires that two similar individuals receive similar treatment. The counterpart, group fairness, requires that demographic subsets of the population are treated equally. These two measurements are not always equivalent, as deciding features may not be evenly spread throughout a population. Within the context of housing, individual fairness states that two prospective customers who differ only in protected traits should be shown the same properties and offered the same rates. Group fairness requires that, on average, a random member of one group should receive the same treatment as a random member of the other.

Throughout this paper we will define bias using the HUD definition of *disparate impact*. This states that "(the) illustrations of unlawful housing discrimination in this part may be established by a practice's discriminatory effect, even if not motivated by discriminatory intent" (78 FR 11459). This is an indirect definition of fairness, as intent and knowledge of protected features are not needed. Therefore, if we conclude that a system is biased we do not attribute any malice to that fact; it is a simple statement that the outcomes or effects are different across demographics.

Housing Bias

Housing discrimination has been a topic of active research since the 1940s (Gaddis 2018), but the introduction of the internet has substantially changed the area. Here we discuss both offline and online audits.

Traditional Housing Audits Housing discrimination in non-internet contexts has been a topic of interest and research for decades. Traditional housing audits are performed in two ways: In the first, activists, nonprofits, or enforcement agencies conduct audits to identify particular discriminatory actors such as racist landlords. In the second, academics or government agencies attempt to characterize discrimination in a large population or area. The first systematic nationwide audit occurred in the UK in 1960 (Gaddis 2018). The US followed with the HUD Housing Discrimination Study (HDS) in 1977, and national HUD studies have been performed in the US in 1989, 2000, and 2012 (Wienk 1979). The 2012 audit showed continued decreases in overall discrimination, but still significant differences in the treatment of prospective renters and buyers along demographic lines.

Online Housing Bias Evidence suggests we should expect bias in housing online. Recent studies of the online rental site Airbnb (Edelman and Luca 2014; Edelman, Luca, and Svirsky 2017) show significant racial bias against those with African American names, while (Edelman, Luca, and Svirsky 2017) demonstrated that African American users were turned down for rentals 58 percent of the time, as opposed to Caucasian renters who were turned down only 50 percent of the time.

Beyond Airbnb, research finds racial bias in online rentals both in the United States (Carpusor and Loges 2006; Hogan and Berry 2011; Hanson and Santas 2014) and in Europe (Ahmed and Hammarstedt 2008). In each of these experiments, researchers emailed landlords to inquire about vacant apartments advertised online and signed the emails with names that suggested the senders' race. Customers with traditionally non-English/American names received fewer responses overall and more negative responses.

Computational Audits

Audits are also conducted on a variety of computational systems. Beyond housing, these audits have covered diverse topics including system privacy, security, advertisements, and search results. Importantly, these audits test both the design of computer systems and the behavior of people that use them.

Information security has presented several unique audit contexts due to the variety and value of the information involved. Previous papers have demonstrated audits for verifying system changes and data provenance from the underlying network topology (Ujcich et al. 2017) all the way up to complex heterogeneous web services (Bates et al. 2017).

Privacy has also been a vital area for computational audits. Gutierrez et al. describe privacy audits for social network sites that would enable users to more easily determine how their data is being shared (Gutierrez et al. 2009). A paper by Bates et al. describes how government-sponsored wiretapping programs can be amended to allow open and transparent auditing from tamper-resistant logs (Bates et al. 2015). Finally, (Oh et al. 2014) describe a system for auditing medical records exchanges while preserving patient privacy, demonstrating how to audit sensitive data sources in a responsible way (Oh et al. 2014).

In addition to auditing systems, researchers can also audit how knowledge of bias can inform user behavior. A 2017 paper by Eslami et al. explored biases in online hotel rating (Eslami et al. 2017). They found that one site's rating system was biased, representing low-to-medium quality hotels as significantly higher in ratings than the authors of the reviews stated. Their study also showed that users modified their behavior in order to compensate for the bias, changing rating techniques and alerting other users of the bias.

Sock-puppet Audits "Sock-puppet" auditing refers to the deployment of automated systems that mimic realistic user behavior in place of the humans used in offline audits. Sock-puppets enable more carefully controlled testing than would be possible with human testers (Sandvig et al. 2014).

Hannák et al. use this audit strategy for detecting differential pricing online (Hannák et al. 2014) using a combination of voluntarily submitted browsing profiles and algorithmically generated browsing behavior. Their experiments showed that more than half of the e-commerce sites audited performed some level of personalization.

The AdFisher paper by Datta, Tschantz, and Datta (Datta, Tschantz, and Datta 2014) describes methods for building user profiles to fit differing demographics by using automated browsing as well as explicit advertisement settings. Their experiments used these sock-puppets to explore a wide range of different topics including transparency and equality across Google search results and ads. They found that women in their study received fewer ads encouraging them to take high paying jobs than men. Although many of the training and measurement techniques described in this paper are no longer feasible due to platform changes, the underlying concepts are foundational to this research area.

The many types of potential bias in online systems have resulted in calls for updated audit methods for online algorithmic systems (Sandvig et al. 2014; Ahmed, Andersson, and Hammarstedt 2010; Chander 2016). Within the domain of housing we have chosen to focus on *advertising bias* and *ranking bias*.

Auditing Advertising Bias As discussed earlier, the contents and presentation of an advertisement can have discriminatory effects. Marketing dangerous or predatory topics to minority audiences can have profound economic, physical, psychological, and societal consequences (Petty, Harris, and Broaddus 2002). Augmenting this with personalized online tracking and targeting may improve the advertisement's efficiency, but it can reinforce existing biases and inequalities by training algorithmic systems with biased data, or for biased outcomes.

For example, Google search results for stereotypically African American names returned ads containing the word "arrest" significantly more than for white names (Sweeney 2013). This creates both individual harms in circumstances such as job searches where employers routinely search for information about the applicant, but also creates societal harms by reinforcing racist stereotypes about African Americans.

A very recent study by Ali et al. showed that Facebook's advertising system can create discriminatory effects without the advertiser's explicit input or awareness (Ali et al. 2019). This presents a dangerous possibility of unknowingly reinforcing negative stereotypes that have been "learned" by a complex system.

Auditing Ranking Bias The influence of search engines has greatly accelerated the study of ranking effects on user interaction. Users overwhelmingly favor the top results in search pages (Guan and Cutrell 2007; Craswell et al. 2008). Keane, O'Brien, and Smyth demonstrated that users continued to click on the top links in Google search results even if the order of the results was reversed (Keane, O'Brien, and Smyth 2008).

Because ordering can effect decision-making so drastically, it is important that the ranking algorithm is fair. A 2018 study analyzed bias in resume search results (Chen et al. 2018). It measured whether there were significant differences between the ranking of male and female job seekers based on inferred gender. The study concluded that the systems were generally fair on an individual basis, but that there was group bias against feminine candidates.

In this work, we investigate advertising and ranking bias in online housing using sock-puppet accounts. This method involves many challenges including profile building and controlling for unprotected attributes that we describe below.

Experimental Design

Our experiments aimed to answer the following questions:

- **RQ1:** Do the profile training techniques presented in this research produce advertising profiles that are significantly differentiated and reflect the desired attributes?
- **RQ2:** Does the perceived race and gender of a user searching for housing online affect the number and/or type of housing advertisements served to that individual?
- **RQ3:** Does the perceived race or gender of a user affect the recommendation ordering on housing listing sites?

We test RQ1 by training profiles with each of the desired classes, browsing the internet to collect ads, and then comparing the contents of the ads served. If a proxy variable for the desired attribute varies significantly between classes we conclude that the training techniques were effective.

Our experimental design for RQ2 and RQ3 drew inspiration from previous audits in the space, especially the nationwide HUD Housing Discrimination Studies (Wienk 1979; Austin Turner et al. 2013). The basic method is a paired test, where each test is performed by at least two individuals who differ only in protected classes. Any significant deviation in treatment between the classes signals an underlying difference in treatment.

Each audit or experiment consisted of hundreds of agents making dozens of measurements. An agent's life cycle consisted of two main steps: 1) training a browser profile and 2) collecting experimental data. Details on the methods of profile training and testing are covered in greater detail in the next section, while data collection is covered in the details of each audit.

Scope

There are a nearly limitless number of audits that could be performed in this area, considering the number of legally protected attributes, size of the United States housing market, and breadth of online services involved in housing. Even performing the simplest audits required limiting several of these variables.

Demographics Each agent in the experiment was assigned an identity at the intersection of two demographic categories: race and gender. All other classes (protected or not) were controlled by removing highly correlated topics.

The first set of categories we considered was gender. Agents were assigned to either the male or female profile training. We recognize that this does not account for transgender or non-binary individuals. However, our lack of data on browsing behavior for these individuals made it prohibitively difficult to define a profile training strategy that included these individuals.

The other categorization was by racial or ethnic background, labeling agents as either Caucasian, African American, Hispanic, or Asian. These were chosen as they are the four largest racial and ethnic groups in the United States. This category is referred to as "race" throughout this paper, although there is active debate as to whether a Hispanic background is racial or ethnic in nature. We chose to use the term race because two-thirds of Hispanic Americans consider their background as part of their racial identity (Center 2015). Additionally, this decision enabled simpler terminology across the audits.

Locations This audit focused on two areas: Chicago, a very large city, and Champaign, a smaller neighboring community.

We include Chicago in this audit due to both size and demographics. It is one of the largest metropolitan areas in the United States, and has a very large number of properties on sale at any time. It is also simultaneously one of the most diverse and most segregated cities in the U.S. (Silver 2015). This diversity increases verisimilitude as it is realistic that people from each of the demographics would be searching for a property in that area. It also presents a situation ripe for geographical discrimination, where landlords or sellers sort buyers into neighborhoods by race.

We included Champaign as a comparison. It is fairly racially diverse, with Hispanic, African American, and Asian populations each comprising at least 10% of the population (Bureau 2018). Furthermore, it is one of the largest communities in the state after Chicago, and therefore has a housing market large enough for a variety of listings in different areas and at different price-points.

Timeline Agents were run in blocks, where a mixture of differing profiles were trained and executed simultaneously. This was done to control for unintended effects that might be caused by changes to the site or network during the course of the experiment.

Blocks of training and collection were scheduled throughout the week and at varying times in order to minimize any data that might be inferred due to the day or time an agent visited a page.

We started collecting results for the advertising audit in early August 2018, and started collecting ranking data in early December 2018. We continued collecting both through mid-March, however we limited the data we analyzed in the Advertising audit to ads collected before February 2019. We chose to limit the data due to the release of several large stories focusing of Facebook's handling of housing ads (Biddle 2019; Robertson 2019). Although these stories did not directly affect our data sources, there was still a possibility that the increased scrutiny in this area would cause the sites we were to testing to change their behavior and invalidate the study.

Technical Details

Each agent is represented as a fresh browser profile at the beginning of the experiment with no cookies or history. Next, it is assigned both a gender and race. It browses the internet following specific patterns in order to build a representative browser profile. Finally, the trained profile visits target sites and collects experimental data for analysis.

An agent was represented by a browser profile in a discrete instance of Firefox 63 running on an Ubuntu 16.04 server. The browser instances were controlled by the Python programming language and the Selenium automation framework. All browsing traffic was routed through a proxy so that it would appear to originate in the correct geographic location.

Profile Training (RQ1)

Because our audit follows the pattern of paired testing, we required a method of providing data about participants (such as age, gender, or race) to the subjects of the audit, online advertising networks and housing sites. This meant building a profile visible to these sites.

These profiles represent prospective home buyers who differ in gender and race, but are similar in location and income. This was done by browsing sites that were disproportionately visited by members of a specific demographic group. This method of inferred profile building was used for two reasons: necessity and verisimilitude.

First, there was no acceptable option for self-selection in the agent's advertising profiles. In earlier papers the experimenters had direct control over interests and visibility into inferred characteristics through tools such as Google Ad Settings (Datta, Tschantz, and Datta 2014; Wills and Tatar

2012). However, using these tools for audits has gotten significantly more difficult in the intervening years. Many sites, including Google, now require an account on the service to access these tools and explicitly forbid creating or using automated accounts in their Terms of Service. Additionally, many advertisers have removed the ability to specify or view the demographic information we need. The combination of these factors made direct profile building infeasible.

Second, this method matches the way that advertisers would determine the interests of a user (Barford et al. 2014). Most networks do not have detailed profile information supplied by the user and must rely on statistical inference. Using browsing to build profiles may reveal latent bias due to inferred data; an advertiser or network may make a decision based on a proxy attribute of a demographic group rather than the demographic itself. This might be missed in an explicit profile, but would be caught with behavioral profiles.

Site Selection

We gathered the representative sites from Quantcast (`quantcast.com`), an internet analytics company that provides demographic breakdowns of site visitors in the United States. The site provides statistics on gender, age, income, ethnicity, and education (Simpson 2019; 2018).

Relative popularity in Quantcast's metrics is measured through the demographic index (Simpson 2019): the probability that a site visitor is a member of that demographic divided by the likelihood that a visitor to a random site is a member of that demographic. Therefore, a demographic index 100 for Hispanic visitors would say that there is no difference in the number of Hispanic visitors from what would be expected on any random site, where a 200 would demonstrate that twice as many visitors are Hispanic than average. For example, the African-American index for ebony.com was 650 in 2019.

When selecting sites, we chose sites that had a demographic index above 140 for gender and ethnic minorities, or above 120 for Caucasian sites. We chose a second threshold for Caucasian users because a small increase in the demographic index for a significant majority group implies a sizable decrease in the number of minority visitors. More specifically, a demographic index for Caucasian visitors of 120 suggests that 90% of the site's visitors are Caucasian, significantly more than the 75% baseline.

In order to create the list of sites, or "playlist", for each demographic, we reviewed the top 500 sites on Quantcast's rankings and selected all of those marked as directly measured by the network that published audience metrics. We then sorted through and collected any sites that matched the demographic index threshold and put them on their corresponding lists.

We removed any sites that had a significantly large demographic index in other areas, specifically age and income. Although age is a protected class, removing sites strongly correlated with a specific age group helps minimize any confounding effects from this unmeasured variable. An agent's income, however, is not a protected attribute. It is important to carefully control income so that it does not confound any other findings.

A set of control sites were also selected from the top 100 most visited sites. These sites had minimal demographic skew, and therefore should not correlate with any specific race or gender. Separate control agents were run using these sites for training alongside the experimental agents to ensure that personalization occurred.

Finally, we verified that each site used Google tracking on their homepages verifying that the Google Ads tracking script was loaded while rendering the homepage. This guaranteed that each page visit would be recorded by the advertising network under test.

During training, each agent visited between twenty and twenty four sites over three sessions. The number of sites visited was limited by the total number of training sites that fit the criteria above. The agent stayed on each site for between one and three minutes, and scrolled around the page in order to roughly mimic human behavior. The exact time and motions were randomized to avoid bot detection. In all, training a profile took roughly one hour.

Profile Validation

One of the main challenges in training profiles was validating them; i.e. whether the assigned gender or race had been correctly inferred by the advertising network. Because there was no direct way to measure this, we used a proxy method to measure profile training accuracy.

The key to this method is finding a feature in the results that is independent of those used to train the profile, and correlated with the desired demographic. For example, if the desired demographic were college-aged students, measuring a significant increase in the number of ads for apartments advertised as "Close to Campus" would indicate that the correct demographic profile had been learned.

This method has several strengths that make it desirable. For one, it is entirely indirect: it does not assume that the advertising system has explicitly labeled the desired demographic value, only that it knows correlations between interest groups, and that the profile fits within the desired set of connected interests.

Additionally, this method can be extended to new advertising networks with relative ease. Direct measurements through tools such as settings pages (Datta, Tschantz, and Datta 2014) require modifying the audit system for every new network. This proxy method can be updated and extended by verifying that the training pages visited and the ads or pages collected are covered by the new network.

Validating Gender in Profile Training In order to validate whether the assigned gender of a profile was accurately inferred, we investigated clothing and jewelry advertisements. These were useful because the majority of these ads were gendered, either by the style of the item or gender of the model. We categorized these advertisements manually, looking up items if the intended gender was unclear. Any items that were not explicitly gendered in the ad or on the seller's website were left uncategorized.

Once all of the apparel related ads were categorized, we counted the numbers of male and female ads for each agent. We collected 167 overtly gendered ads, of which 110 were gendered female. We then performed a paired t-test which showed a statistically significant correlation ($p = 0.046$) between the number of female-gendered ads seen and the gender of the agent. This supported our hypothesis that the advertising network learned the intended feature.

Validating Race in Profile Training While we could not validate all categories of races, we used Spanish vs. English language ads to verify the Hispanic training. We categorized Spanish language ads and compared the frequency of these ads between Hispanic and non-Hispanic profiles. We found that while Hispanic profiles received some Spanish ads (63 in total), non-Hispanic profiles never received a Spanish ad. While not comprehensive, this suggests that our profile training method has distinguished between Hispanic and non-Hispanic profiles.

This method of profile training and verification is very flexible. Because there is no aspect unique to housing, it can be extended to other sock-puppet audits in order to measure the effects of both gender and race.

The Advertising Audit (RQ2)

The first audit concerns housing-related ads served to users while browsing the internet. Our goal was to determine whether the number of housing related ads served to an agent was correlated with their gender or race.

Data Collection

Once trained, the sock-puppet profiles browsed the internet and collected the ads served to them. Three main categories of pages were chosen, distinct from the training sites, on which to collect ads: Google search results, local newspapers' online housing listings, and national news homepages. Search results were chosen as a source of data due to the high proportion of relevant ads. A collection of different search terms were used by each agent, focusing on houses for sale or rent in the surrounding area. These terms included "houses for sale in $location", "$location rentals", and "houses in $location". These search terms were chosen to portray an interest in housing related topics as well as re-enforcing the desired target location. Local newspapers were chosen as a category due to the relevant page topics and high number of display advertising slots served. The national news pages were chosen as they target a neutral audience, and because they provide a good baseline for the ads an agent would see on sites that are not directly related to housing topics.

While selecting sites within each of the categories above, we verified that each site served ads through Google's AdSense network. This, when combined with the requirement training pages include Google tracking, ensured that an agent's profile could be correlated across throughout both the training and data collection phases.

In addition to the text, image, or frame-grab of an ad, we collected relevant metadata including information about the search term used, time of day, target URL, and the agent that collected the ad. We did not collect the position of each ad because the framework code that retrieves the location of an element returned erroneous results.

Ad Categorization

The next step after visiting the aforementioned websites was finding and categorizing the housing ads.

Although the question of how many overall housing-related ads were seen is important, it gives very little information on how the users' treatment differed. It was important to have more detailed descriptions of the content of advertisement in order to better understand the mechanics of the advertising networks and investigate possible explanations for behavior.

To this end we defined categories of relevant ads, focusing on sellers and their relationship to the user. This was done so that we could determine if any particular class of advertiser displayed bias, rather than measuring the entire ecosystem at once.

Housing-related ads were placed into one of the following categories:

- **Sale Listing:** An ad to purchase a specific property currently on the market, stating either the address or community name.
- **Rental Listing:** As above, but an ad to rent. This includes short-term rentals.
- **Realtor:** An advertisement for a real-estate agent, group, or property manager. These ads must be directly associated with a professional who can broker a property sale or rental.
- **Rental Group:** An ad for a company that facilitates rentals, but does not own or directly manage the properties.
- **Foreclosure:** Ads for search engines and sellers specializing in foreclosed properties, including the words "foreclosed" or "bank-owned".
- **Rent-to-Own:** Any ad that used the term "rent-to-own" was included in this category, unless it contained a listing that was currently for sale.
- **Listing Aggregator:** A search engine or site that collects other listings.
- **Property Search Engine:** A service that collects and searches property information, but is not primarily intended for buying/selling.
- **Loans:** Ads that directly reference housing-related loans such as mortgage services. Excludes credit card ads, savings accounts, or other general banking ads.

The number of important categories, the subtle differences in requirements, and inclusion of image and video ads made manual classification the most sensible method. Categorization was performed in batches by one of the authors, working from a list of unique advertisements collected.

Results

We collected 21,740 ads using 766 trained profile agents each assigned to one of nine distinct treatments — roughly 95 of each of the eight experimental treatments and 12 of the control. Of these ads, 2,262 were unique.

For a given category, we compared the number of ads observed by each group using a two-way ANOVA test. Using

	Female	Male
African American	15.17	16.74
Asian	16.38	16.13
Caucasian	16.15	17.97
Hispanic	15.65	15.92

Table 1: The average number of housing-related ads seen by agents of each group. This audit collected 12,418 housing-related ads in total over 766 profiles.

	Df	F val	P val
Race	3	2.30	0.057*
Gender	1	2.48	0.116
Race × Gender	3	1.17	0.319

Table 2: The results of a two-way ANOVA test with the number of ads seen as dependent on the agent's gender and race. The p-value for race is not quite lower than the standard 0.05 level, but suggests that there may be some differential treatment, even if it is not statistically significant. n = 12,418

this approach, we could test both demographic variables in a single experiment.

The ANOVA test requires the data be both normal and homogeneous in order to make accurate conclusions; this was verified before the statistical analysis. We verified these assumptions using the Fligner-Killeen test for homogeneity of variance, the Shapiro-Wilk normality test, and inspection of both the Residual vs. Fitted and Normal Q-Q plots.

Housing-Related Ads We first investigated whether there was any bias in the number of general housing-related ads served. This included all of the categories listed above except for unrelated ads. Of the 21,740 total ads, 12,418 were housing-related, with the average number seen per-demographic in table 1.

In table 2, we can see that the ANOVA test detects some correlation ($p = 0.0569$) between race and the number of housing-related ads served. Although this is not conclusive evidence of bias at the $p = 0.05$ level, it suggests the ads were not evenly distributed.

Inspecting table 1 we see that the average number of housing-related ads seen was generally clustered between 15.50 and 16.50, except for Caucasian men who saw more ads than average, and African American women who saw fewer.

In order to investigate this further, we analyzed the data pair-wise: selecting a subset of the dataset and comparing it to its complement. Running the ANOVA test comparing Caucasian users with non-Caucasian users shows a significant difference ($p = 0.033$) in treatment, while the other three permutations show significance values close to the non-pairwise test ($p = 0.060 \pm 0.005$).

This demonstrates that, while there may not be significant differences when considering all variables simultaneously, the treatment of Caucasian agents was significantly different than the treatment of others. Since the data for Caucasian visitors never made up more than one third of the data under consideration, this bias was not as clear to the original method of analysis.

	Df	F val	P val
Race	3	0.79	0.498
Gender	1	0.34	0.561
Race × Gender	3	0.11	0.956

Table 3: The results of a two-way ANOVA test with the agent's gender and race as independent variables and the number of listing advertisements as the dependent variable. This test does not show any evidence of bias on any of the demographics. n = 444

	Female	Male
African-American	50	70
Asian	53	50
Caucasian	50	33
Hispanic	60	48

Table 4: The total number of rent-to-own and foreclosure ads seen by each demographic.

Listing Ads The next category of ads we surveyed was property listings. This is an important category to consider as it is the most directly regulated by the FHA.

We do not see any significant bias in the number of listings served using the multi-factor ANOVA test (see table 3). Unlike the previous category, there is no evidence of bias found by pairwise tests either. This is an encouraging result because it does not demonstrate underlying bias. However, we cannot conclude that the system is unbiased in this area without further analysis. This will be addressed further in the Discussion section.

Rent-to-Own Ads One area of housing loans that is particularly interesting is that of rent-to-own properties. Also known as "contract-for-deed" arrangement, they promise the deed to a property (or the option to purchase the deed) upon completion of contract terms, usually including a set period of on-time payments and paying both taxes and upkeep for the property.

There has been significant discussion about the conditions of these agreements, with the various state investigators and the US Department of Defense describing them as a form of predatory lending (Goldstein and Stevenson 2016; Department Of Defense Washington Dc 2006). They are generally targeted at low-income buyers, and previous lawsuits have alleged that these companies specifically target African American buyers and neighborhoods (Oakland vs. Wells Fargo).

A pairwise analysis shows that African-American users received more ads for this type of loan than any other race ($p = 0.017$). Additionally, in table 4 we can see that Caucasian men saw far fewer ads of this type than any other combination of race and gender. This supports previous assertions that these subprime loans are being targeted at minority communities.

Other Categories of Ads We did not see any significant differences in treatment for any other categories of housing ads. This was due to a few limiting factors. First of all, many of the categories were seen very infrequently, which limits the statistical power of our techniques. Secondly, several categories had a bimodal distribution in the number of views. In these, a few agents saw ads many times, but most agents never saw any. This violates many of the assumptions of the ANOVA test, and also compounded the first limitation. Finally, a few of the categories were sufficiently large, but failed one or more of the requirements of our statistical tests.

The first and third limitations may be addressed in the future with significantly larger sample sizes, under the assumption that the number of ads seen roughly follows a normal distribution. The second, however, may be a problem regardless of the sample size if that pattern does not hold. Thus, alternative analysis modeling the number of ads as a negative binomial or Poisson distribution may be more applicable for situations where this occurs.

The Ranking Audit (RQ3)

The other audit we performed was an analysis of whether the gender or race of a user is correlated with the ranking of recommended property search results. This analysis focused specifically on online listing services.

In order to select sites for this audit we performed an initial survey of several of the most popular online listing services, including Zillow, Trulia, Redfin, Realtor.com, Homesnap, and Forsalebyowner [1]. Sites were considered for inclusion if they met a series of criteria: First, it must have an algorthmically generated "recomended" sort. This immediately ruled out a few sites that only sorted based on user-visible characteristics. Second, that sort must be personalized. Both Zillow and Redfin were removed as they did not exhibit any personalization. Both display pre-generated recommendations that were identical for every agent that visited the site. Finally, we chose sites that allowed scraping the pages of interest in their robots.txt files. This was to done in order to be "good neighbors" to these sites by only scraping pages they allow.

This left us with two popular listing sites to audit: Trulia and Realtor.com. Both allow searching properties for rental or sale and provide a personalized sort option. Additionally, both include Schema.org metadata in their search pages. This simplified data collection and reduced the number of requests to the respective sites.

Data Collection

In the data collection process, the trained browser profile visited the homepage and submitted the desired location using the main search bar. When the results were returned, the program verified that they were sorted using the personalized sort. On Trulia this is called "Just for you", and "Relevant Listings" on Realtor.com. If they were not sorted correctly, the desired sort was selected from the drop-down box. The program then scrolled down the page while extracting the metadata for each listing from the page's HTML. We collected the unique ids of the listings along with their addresses, latitude and longitude, number of bedrooms and bathrooms, and price.

Our study focused on single-family residences. Commercial properties and larger properties consisting of multiple units for sub-letting were not considered in our audits. Before removing these properties from the dataset, we verified that they were seen at the same frequency and positions for visitors of all demographics.

Pricing Results

We used a similar technique to analyze the rankings as we did the advertisements. For each listing service, we performed a multi-factor ANOVA test with the rank position as the independent variable, the profile variables (gender and race) as the control variables, and the price as the dependent variable.

Statistically significant interaction effects in the ANOVA test suggest that the listings are weighted differently depending on the variables involved. For example, a significant result for race would suggest that at least one group was shown properties that were substantially cheaper or more expensive than those shown to others. Below, we describe these analyses and results in detail.

We also verified that personalization only influenced the presentation (ordering and visibility) of listings, and not the attributes (price, square footage, etc…) of the listings themselves. Most importantly, our audit did not see differential pricing on any site we surveyed. This is a welcome departure from the 2012 HUD audit that found substantial price discrimination in one-on-one interactions (Austin Turner et al. 2013).

Realtor.com We collected 436 measurements from Realtor.com over the course of three months. Each of these records contains the metadata for each of the top 30 listings as well as the time collected, unique identifier of the browser profile, and the trained gender and race of the profile. In total we collected information on 789 distinct properties.

Next we performed a multi-factor ANOVA test on the provided data. The results of the test are in table 5. We see that there is a statistically significant ($p = 0.022$) interaction effect between gender and position in the results. This implies that men and women receive different orderings. Looking at the plot of the ANOVA coefficients for gender, position, and price in fig. 1, we see that women are presented more expensive properties than men at the top of the list, but less expensive properties as they move down the list.

We also surveyed the listings displayed by price alone. Table 6 lists the average listing price seen in each location by race. Female agents in Chicago receive recommendations for more expensive properties than their male counterparts, except for Asian users. However, this pattern re-

[1] zillow.com, trulia.com, redfin.com, realtor.com, homesnap.com, forsalebyowner.com

	Df	F val	P val
Race	3	0.20	0.898
Gender	1	1.08	0.298
Position	1	0.33	0.566
R × G	3	0.23	0.877
R × P	3	0.12	0.947
G × P	1	5.24	0.022**
R × G × P	3	0.44	0.724

Table 5: A three-way ANOVA test on Realtor.com. We can see a significant interaction between gender and position, implying that the rankings are ordered differently by price depending on the gender of the user.

	Chicago		Champaign	
	Female	Male	Female	Male
AA	$378,123	$343,619	$224,922	$224,173
A	$338,621	$364,871	$226,059	$220,709
C	$395,231	$317,609	$220,961	$227,590
H	$359,625	$344,918	$217,756	$230,006

Table 6: The average price (in USD) of listings seen on Realtor.com. 2961 records were collected from Chicago, and 2474 from Champaign (Place names anonymized)

verses in Champaign. There, men receive recommendations for slightly more expensive properties than women. Once again, the trend is swapped for Asian agents, with Asian women receiving more expensive recommendations. Note that this is not statistically significant. More analysis is necessary to understand the associations between these agents and the recommended properties.

Trulia.com In the Trulia audit we collected 846 measurements over the course of three months. Like Realtor.com, each measurement consisted of the first 30 recommended properties as well as metadata about the scraping profile.

The results of the multi-factor ANOVA test do not show any evidence of bias towards either of the profile variables or position. Similarly, we did not see any interaction effects in the average price of listings shown to any demographic. The average prices for all demographics at any position fall within ±$10,000, substantially less than any of those for Realtor.com.

We verified that these effects were not due to the larger sample size of Trulia results by randomly sampling a subset of the results and calculating the averages of the subset. When tested, these samples shared the same small deviations as the complete dataset. This suggests that the degree of personalization in the price of properties shown is smaller on Trulia than Realtor.com for the profile variables in this audit. Further research would be required to prove this conclusively.

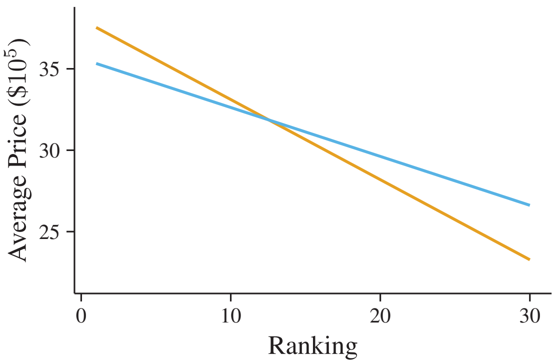

Figure 1: Comparison of the ANOVA coefficients for each gender at a given index, with the female line in orange and the male line in blue. While the trend-line of the average price for women starts above the one for men, the two cross halfway down. This implies that, although the mean price is similar, women are recommended more expensive properties first.

	AA	A	C	H	P val
Lake View	8	15	19	21	0.099*
Loop	8	7	9	8	0.969
Norwood Park	5	4	6	5	0.940
Cicero	8	10	5	5	0.630
Chatham	9	11	11	10	0.966

Table 7: The number of times a neighborhood was seen by each racial group in Realtor.com results. The Lake View neighborhood was recommended to African American users fewer times ($p = 0.099$) at the $p = 0.10$ significance level.

Geographical Discrimination Results

Another area of interest for this audit was whether the race of a user influenced the neighborhoods they were recommended. This leads to indirect segregation: members of minority groups are technically allowed to move outside their neighborhoods, but are prevented from doing so because they are unable to find funding. This results in racially segregated communities as well as depriving minority groups from building greater net worth (Henkel, Dovidio, and Gaertner 2006; Pincus 1996).

This portion of the audit uses the same set of data as the other ranking experiments, but a different statistical test. Because we are interested in how many users are recommended each neighborhood, we use a simple χ^2-test rather than ANOVA.

We saw very little variance in the number of times a neighborhood was seen by each racial group on Trulia. The least evenly served neighborhood was Douglas, with $p = 0.7468$. This is well above our threshold of significance, and therefore we conclude that we did not see any evidence of geographical discrimination on Trulia.

We see a different story on Realtor.com. Looking at the values in table 7, we see that there is one neighborhood, Lake View, with $p = 0.099$. Further restricting to the top 15 listings in the ranking decreases this value further to $p = 0.059$. While still above the standard $p = 0.05$ confidence threshold, this suggests that the neighborhood is not only shown less to African American users, but when it is shown, it is placed further down as well.

This difference is especially compelling when considering the population of the Lake View neighborhood. This neighborhood is in the top 10% of the city by median income, and is also drastically less diverse than the city as a whole. While Chicago is 45% White and nearly 35% African American (for Planning 2018a), Lake View is almost 80% White and less than 5% African American (for Planning 2018b). If systematic, a significant difference in recommendations by race combined with the population dynamics of the neighborhood could result in de facto segregation.

Discussion

In this paper we described two audits on online housing markets; the first studying online ad targeting and the second measuring the effects of personalized search result ranking.

In the first audit, we found differences in the number of housing ads shown on the basis of race, and demonstrated that Caucasian agents saw significantly more housing-related advertisements ($p = 0.0569$ overall, $p = 0.033$ pairwise) than other agents who searched for housing. This suggests that the advertiser's target audiences were more closely aligned with the Caucasian group, leading to indirect bias. The first audit also found that advertisements for predatory rent-to-own programs were seen by African American users significantly more ($p = 0.007$) than any other group.

The second audit uncovered evidence of gender bias ($p = 0.022$) in the ordering of suggested properties on the online housing site Realtor.com. Women were recommended more expensive properties towards the top of the list, and less expensive properties towards the bottom. We also concluded that African American users were somewhat less likely to be recommended houses in the Lake View neighborhood of Chicago($p = 0.099$). While this may be unintentional, techniques like this have been used in the past to enforce indirect segregation by pricing minorities out of desirable neighborhoods (King and Mieszkowski 1973).

While our results show some bias in these systems, we do not claim that any laws have been broken. Further audits would be required to show a pattern of systematic bias in one of these platforms.

Auditing Framework

The audits presented here demonstrate a framework for extending traditional paired-testing methods to automated online systems. The techniques developed for profile training can be extended to other demographic groups by developing new site playlists and selecting suitable verification variables. Researchers studying advertisements in a geographical area can select sites with local interest. This would be especially relevant where the site's readership is correlated with the demographics studied, as differences in the ads shown site-by-site could result in bias. Finally, additional locations, sites, and property attributes could be added to further studies.

Concluding an Absence of Bias

While we can observe differences in treatment from the two audits, our current statistical methods cannot conclude that there is no bias in these systems. This is due to the scientific difficulty of proving a negative, and follows the historical focus on the presence of bias rather than its absence. In further audits, researchers could leverage Bayesian techniques or equivalence testing to prove that these systems treated users fairly.

A 2017 paper by Thebault-Spieker et al. demonstrated some of these methods for proving the absence of bias in online 5-star ranking systems (Thebault-Spieker et al. 2017). While their methods are not directly applicable to housing audits, similar ideas could be applied to the housing space.

Reproducibility

One weakness of these audits is a lack of reproducibility. Practically every factor in these systems is constantly evolving, from the set of ads currently being served, to the targeting and pricing of an advertising campaign, and even the way user profiles are interpreted. This puts researchers in a difficult position: auditors must collect as much data as possible in order to catch any confounding variables, and must carefully validate that the system they are measuring did not change substantially during the course of their audit.

Researchers looking to replicate the findings of an audit like these can use the same tools and methods as the original experiment, but should be very aware that any number of variables may have changed between their measurements and the originals, and therefore could lead to different results. This presents a very compelling argument for designing audits whose practices and results can be independently verified, as well as sharing the collected datasets. The data collected for this audit, including the images of ads collected, is available along with the code at (Asplund 2020).

Detecting and Resolving Bias

Many sites address the liability caused by personalized recommendations by serving an algorithmically sorted list to all visitors and refreshed regularly. This guarantees that two users with the same query at the same time will receive the same recommendations. For areas where fairness is legally required, such as housing and employment, this can be a prudent alternative. The algorithm may still perpetuate existing bias by under- or over-representing certain neighborhoods and communities, but the system as a whole would be fair under this style of paired-testing audit.

If a service provider chooses to serve targeted housing ads or listings, it is important that they regularly audit their own systems for fairness. This is especially important if the service uses opaque or uninterpretable decision-making systems. These would preferably be performed by an unbiased third-party, but could also be accomplished by an isolated team within the company.

Finally, companies should open up access to their systems for fully independent auditing. As it currently stands, external audits not officially sanctioned by the company run the risk of violating the site's Terms of Service, especially if the researchers access the site using automated tools. As long as these sites are not open to external experiments, the public has no way to verify that they meet the legal standard of fairness.

Conclusion

Our two audits investigate important emerging areas of housing fairness: online advertising and search-result ranking. Within the context of advertising, we observed that Caucasian users saw substantially more housing-related ads compared to other races. Additionally, African American men saw significantly more advertisements for controversial rent-to-own advertisements. In the ranking audit we showed that the ordering of properties on Realtor.com was effected by the gender of the user, placing more expensive properties at the top for female users. Finally, we saw some evidence of geographic discrimination where the Lake View neighborhood was seen less often by African American users.

This proposed audit framework can provide a foundation for further audits in the future. The most recent Housing Discrimination Study surveyed 28 metropolitan areas across the United States of America in order to provide a conservative measure of housing discrimination in the county. An online audit of this scale would certainly be a sizeable undertaking, but the lower requirements for participants, travel, and other logistical concerns could make it feasible to conduct on a more regular basis than the roughly ten year pattern of existing audits.

The profile building technique we have used in our auditing platform is also adaptable. An auditor can add new categories or variables simply by changing the sites visited, the training schedule, or the agent's behavior while browsing a page. Even attributes such as the IP address and browser used can be modified in order to build a more representative profile. These behavioral models could be further strengthened by user studies providing truly representative browsing data.

Acknowledgements

We thank the National Science Foundation for their support of this work (NSF CHS-1564041 & CHS-1564079).

References

Ahmed, A. M., and Hammarstedt, M. 2008. Discrimination in the rental housing market: A field experiment on the Internet. *Journal of Urban Economics* 64(2):362–372.

Ahmed, A. M.; Andersson, L.; and Hammarstedt, M. 2010. Can Discrimination in the Housing Market Be Reduced by Increasing the Information about the Applicants? *Land Economics* 86(1):79–90.

Ali, M.; Sapiezynski, P.; Bogen, M.; Korolova, A.; Mislove, A.; and Rieke, A. 2019. Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. *arXiv:1904.02095 [cs]*. arXiv: 1904.02095.

Asplund, J. 2020. Data for: Auditing race and gender discrimination in online housing markets. https://doi.org/10.13012/B2IDB-1408573 V1

Austin Turner, M.; Rob, S.; Levy Diane, K.; Wissoker, D.; Aranda, C.; and Pitingolo, R. 2013. Housing discrimination against racial and ethnic minorities 2012. *Washington, DC: US Department of Housing and Urban Development*.

Barford, P.; Canadi, I.; Krushevskaja, D.; Ma, Q.; and Muthukrishnan, S. April 2014. Adscape: Harvesting and analyzing online display ads. In *Proceedings of the 23rd International Conference on World Wide Web*, 597–608.

Bates, A.; Butler, K. R. B.; Sherr, M.; Shields, C.; Traynor, P.; and Wallach, D. 2015. Accountable wiretapping – or – I know they can hear you now. *Journal of Computer Security* 23(2).

Bates, A.; Hassan, W. U.; Butler, K.; Dobra, A.; Reaves, B.; Cable, P.; Moyer, T.; and Schear, N. 2017. Transparent Web Service Auditing via Network Provenance Functions.

Biddle, S. 2019. Facebook's Ad Algorithm Is a Race and Gender Stereotyping Machine, New Study Suggests, https://theintercept.com/2019/04/03/ facebook-ad-algorithm-race-gender/.

Bureau, U. C. 2018. U.S. Census Bureau QuickFacts: Champaign County, Illinois, https://www.census.gov/quickfacts/champaigncountyillinois.

Carpusor, A. G., and Loges, W. E. 2006. Rental Discrimination and Ethnicity in Names. *Journal of Applied Social Psychology* 36(4):934–952.

Center, P. R. 2015. Multiracial in America: Proud, diverse and growing in numbers. *Pew Research Center Social & Demographic Trends*.

Chander, A. 2016. The Racist Algorithm 2017 Survey of Books Related to the Law: Reviews. *Michigan Law Review* 115:1023–1046.

Chen, L.; Ma, R.; Hannák, A.; and Wilson, C. 2018. Investigating the Impact of Gender on Rank in Resume Search Engines. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*.

Collins, H., and Khaitan, T. 2018. *Foundations of Indirect Discrimination Law*. Bloomsbury Publishing.

Craswell, N.; Zoeter, O.; Taylor, M.; and Ramsey, B. 2008. An Experimental Comparison of Click Position-bias Models. WSDM '08, 87–94. ACM.

Datta, A.; Fredrikson, M.; Ko, G.; Mardziel, P.; and Sen, S. 2017. Proxy Non-Discrimination in Data-Driven Systems. *arXiv:1707.08120 [cs]*.

Datta, A.; Tschantz, M. C.; and Datta, A. 2014. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. *arXiv:1408.6491 [cs]*.

Department Of Defense Washington Dc. 2006. Report on Predatory Lending Practices Directed at Members of the

Armed Forces and Their Dependents:. Technical report, Defense Technical Information Center.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*.

Edelman, B. G., and Luca, M. 2014. Digital Discrimination: The Case of Airbnb.com. SSRN Scholarly Paper ID 2377353, Social Science Research Network, Rochester, NY.

Edelman, B.; Luca, M.; and Svirsky, D. 2017. Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. *American Economic Journal: Applied Economics* 9(2):1–22.

Eslami, M.; Vaccaro, K.; Karahalios, K.; and Hamilton, K. 2017. "Be Careful; Things Can Be Worse than They Appear": Understanding Biased Algorithms and Users' Behavior Around Them in Rating Platforms. In *Eleventh International AAAI Conference on Web and Social Media*.

for Planning, C. M. A. 2018a. Community Data Snapshot | Chicago, https://www.cmap.illinois.gov/documents/10180/102881/Chicago.pdf.

for Planning, C. M. A. 2018b. Community Data Snapshot | Lakeview, https://www.cmap.illinois.gov/documents/10180/126764/Lake+View.pdf.

Gaddis, S. M. 2018. *Audit studies: Behind the scenes with theory, method, and nuance*, volume 14. Springer.

Goldstein, M., and Stevenson, A. 2016. 'Contract for Deed' Lending Gets Federal Scrutiny. *The New York Times*.

Guan, Z., and Cutrell, E. 2007. An Eye Tracking Study of the Effect of Target Rank on Web Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Gutierrez, A.; Godiyal, A.; Stockton, M.; LeMay, M.; Gunter, C. A.; and Campbell, R. H. 2009. Sh@re: Negotiated audit in social networks. In *2009 IEEE International Conference on Systems, Man and Cybernetics*, 74–79.

Hannák, A.; Soeller, G.; Lazer, D.; Mislove, A.; and Wilson, C. 2014. Measuring Price Discrimination and Steering on E-commerce Web Sites. IMC '14, 305–318. ACM.

Hanson, A., and Santas, M. 2014. Field Experiment Tests for Discrimination against Hispanics in the U.S. Rental Housing Market. *Southern Economic Journal* 81.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. *arXiv:1610.02413 [cs]*.

Henkel, K. E.; Dovidio, J. F.; and Gaertner, S. L. 2006. Institutional Discrimination, Individual Racism, and Hurricane Katrina. *Analyses of Social Issues and Public Policy* 6(1):99–124.

Hogan, B., and Berry, B. 2011. Racial and Ethnic Biases in Rental Housing: An Audit Study of Online Apartment Listings. *City & Community* 10(4):351–372.

Keane, M. T.; O'Brien, M.; and Smyth, B. 2008. Are People Biased in Their Use of Search Engines? *Commun. ACM* 51(2):49–52.

King, A. T., and Mieszkowski, P. 1973. Racial Discrimination, Segregation, and the Price of Housing. *Journal of Political Economy* 81(3):590–606.

Narayanan, A. 2018. 21 fairness definitions and their politics. *New York, NY, USA*.

Oh, S. E.; Chun, J. Y.; Jia, L.; Garg, D.; Gunter, C. A.; and Datta, A. 2014. Privacy-preserving Audit for Broker-based Health Information Exchange. CODASPY '14.

Pedreschi, D.; Ruggieri, S.; and Turini, F. 2008. Discrimination-aware data mining. In Li, Y.; Liu, B.; and Sarawagi, S., eds., *Proceedings of the 14th ACM SIGKDD*.

Petty, R. D.; Harris, A.-M. G.; and Broaddus, T. 2002. Regulating Target Marketing and Other Race-Based Advertising Practices. *Mich. J. Race & L.* 8:61.

Pincus, F. L. 1996. Discrimination Comes in Many Forms: Individual, Institutional, and Structural. *American Behavioral Scientist* 40(2):186–194.

Robertson, A. 2019. What happens next in the housing discrimination case against Facebook? *The Verge*.

Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. *Data and discrimination: converting critical concerns into productive inquiry*.

Schwartz, A. F. 2014. *Housing Policy in the United States*. Routledge.

Silver, N. 2015. The most diverse cities are often the most segregated, https://fivethirtyeight.com/features/the-most-diverse-cities-are-often-the-most-segregated/.

Simpson, S. 2018. Reading Our Audience Measurement Reports, http://help.quantcast.com/hc/en-us/articles/115014120368-Reading-Our-Audience-Measurement-Reports.

Simpson, S. 2019. Glossary, http://help.quantcast.com/hc/en-us/articles/115013851427-Glossary.

Sweeney, L. 2013. Discrimination in Online Ad Delivery. *Queue* 11(3):10:10–10:29.

Thebault-Spieker, J.; Kluver, D.; Klein, M. A.; Halfaker, A.; Hecht, B.; Terveen, L.; and Konstan, J. A. 2017. Simulation Experiments on (the Absence of) Ratings Bias in Reputation Systems. *Proceedings of the ACM on Human-Computer Interaction* 1(CSCW):1–25.

Ujcich, B. E.; Miller, A.; Bates, A.; and Sanders, W. H. 2017. Towards an accountable software-defined networking architecture. In *2017 IEEE Conference on Network Softwarization (NetSoft)*, 1–5.

Wienk, R. E. 1979. *Measuring racial discrimination in American housing markets: the housing market practices survey*.

Wills, C. E., and Tatar, C. 2012. Understanding What They Do with What They Know. In *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society*.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning Fair Representations. In *International Conference on Machine Learning*, 325–333.