Discrimination and Design: Equity in Digital Platforms

Michael Luca, *Johns Hopkins Carey Business School*
Ben Workman, *MIT*

Abstract

Digital platforms mediate much of modern economic and social exchange, reshaping how bias operates and how it can be detected, mitigated, or amplified. This chapter develops a conceptual framework linking economic models of discrimination—including taste-based, inference-based, and attention-based—to platform and algorithm design. We synthesize empirical evidence, present a stylized model showing how voluntary interventions can be improved or backfire through selection, and organize design responses into four categories: information provision, choice architecture, changes to algorithms, and regulation. We also review methods for inferring user race and their trade-offs. The framework clarifies how design choices shape and reveal bias, offering a foundation for empirical and policy analysis of equity in digital platforms.

1. Introduction

Disparities between demographic groups remain stark across many domains. A large body of evidence has documented discrimination against minority groups in labor markets (Pager 2003; Bertrand and Mullainathan 2004; Quillian et al. 2017; Kline et al. 2022), credit markets (Munnell et al. 1996; Bartlett et al. 2022; Butler et al. 2022), and political elections (Stephens-Davidowitz 2014; Broockman and Soltas 2020).

As a growing share of our interactions move online and into digital environments, platforms have become central to how people find jobs, housing, and services. Discrimination remains common across online settings, including labor platforms (Acquisti and Fong 2020; Chan 2024), housing (Edelman et al. 2017; Laouénan and Rathelot 2022), peer-to-peer marketplaces (Doleac and Stein 2013; Ayres et al. 2015), and transportation (Ge et al. 2020). In principle, platform design choices could help mitigate bias—for instance, by altering algorithms, anonymizing user attributes, or changing the choice environment. In other cases, research has documented evidence of support for historically disadvantaged groups. Here too, platform design choices can play a role.

This chapter examines evidence of discrimination on platforms, how it connects to economic models of bias, and how design choices may shape outcomes. When race or other demographic attributes are made salient, discrimination has often been observed in platform-mediated transactions. Some platform designs, such as Airbnb's use of profile photos, make race more visible. By contrast, other platforms make race less salient, though signals such as names, language, or images can still convey demographic information and engender discrimination. These patterns can be interpreted through the lens of economic models of discrimination. In addition to the foundational models of taste-based discrimination (Becker 1957) and statistical discrimination (Phelps 1972; Arrow 1973), patterns of bias on platforms may be better explained by models that incorporate mechanisms such as biased beliefs and limited attention.

Consider, for example, evidence suggesting that online ratings can reduce the extent of bias in digital marketplaces. One explanation is that ratings reveal individual quality, reducing reliance on group-based proxies as in models of statistical discrimination. However, ratings may also affect behavior by channeling users' attention toward alternative features or updating beliefs that may otherwise be biased. In Chan (2024), racial disparities in willingness to pay for physicians shrink dramatically when quality ratings are available. While the results are directionally consistent with statistical discrimination, the magnitude of the effect suggests that additional mechanisms—such as belief distortion or limited attention—may also be important.

Insights into when and why bias arises on platforms can also help guide the design of platform interventions. We consider ways in which interventions—such as concealing or revealing demographic information—can reduce or amplify discrimination. The impact of these choices can depend on factors such as the mechanism driving discrimination and the selection of users who are exposed to the information. More broadly, choice architecture and algorithmic design shape user behavior and interact with the underlying sources of discrimination.

While economic theory helps inform platform design—by identifying mechanisms that give rise to bias—platforms also serve as real-world environments for testing these models. They offer natural variation, experimental opportunities, and high-frequency behavioral data, making them fertile ground for evaluating when and why discrimination emerges.

The remainder of the chapter is organized as follows. Section 2 presents empirical evidence on bias and connects it to models of discrimination developed in the economics literature. Section 3 discusses efforts platforms have made to support marginalized groups, both by reducing discrimination and by explicitly channeling support for these groups. Section 4 reviews possible ways to measure race on platforms and discusses the tradeoffs involved with each. Section 5 concludes.

2. Discrimination on Digital Platforms: Evidence and Theory

2.1 Empirical Evidence on Discrimination in Online Platforms

A growing empirical literature documents how discrimination manifests across digital platforms. Table 1 presents a set of illustrative findings drawn from studies on platforms ranging from housing and transportation to retail, healthcare, and labor markets. While many focus on racial disparities, others identify discrimination based on religion (Acquisti and Fong 2020) or disability status (Ameri et al. 2020). Platforms vary in how prominently demographic signals are surfaced—through features like names, profile photos, or user-facing identity fields—reflecting both platform design and user norms. These design choices likely influence the extent of discrimination in routine user interactions. Together, these examples highlight the range of contexts in which bias has been measured and provide a starting point for connecting platform-specific evidence to broader economic theories of discrimination.

Table 1: Evidence of Bias on Digital Platforms

Paper	**Platform**	**Main Finding**
Edelman et al. (2017)	Airbnb	Applications from guests with distinctively African American names were 16% less likely to be accepted than those with distinctively White names.
Ameri et al. (2020)	Airbnb	Hosts were more than twice as likely to reject booking requests from travelers with disabilities like blindness, cerebral palsy, and spinal cord injury.
Laouénan and Rathelot (2022)	Airbnb	Ethnic minority hosts charged 3.2% less than White hosts of comparable listings.
Luca et al. (2024)	Airbnb	Hosts with distinctively Asian names on Airbnb saw a 20% decline in bookings compared to hosts with distinctively White names at the start of COVID-19, a time of heightened bias against Asian Americans.
Ge et al. (2020)	Uber	The cancellation rate was twice as high for passengers with distinctively African American names relative to passengers with distinctively White names.
Ayres et al. (2015)	eBay	In an auction, baseball cards held by a Black hand sold for 20% less than those held by a White hand.
Doleac and Stein (2013)	Craigslist	iPods held by a Black hand received fewer and lower offers than iPods held by a White hand, especially in less competitive markets.
Acquisti and Fong (2020)	LinkedIn	Employers in more Republican areas called back Muslim candidates less frequently than Christian candidates.
Evsyukova et al. (2024)	LinkedIn	Profiles with Black profile pictures were 13% less likely to have connection requests accepted relative to those with White profile pictures.
Chan (2024)	Online healthcare bidding marketplace	In an online healthcare bidding marketplace, customers were willing to pay up to 12.7% less for Black doctors and 8.7% less for Asian doctors than

		for White doctors.

Note: This table presents illustrative examples from empirical studies documenting discrimination on digital platforms.

2.2 Connection to Theories of Discrimination

The papers presented in the previous section document discrimination in a variety of settings on digital platforms. Understanding *why* this discrimination occurs is important both for guiding interventions designed to reduce bias and for advancing our understanding of discrimination. In this section, we briefly review prominent models of discrimination in the economics literature as they relate to the empirical evidence.

Economics research has focused heavily on two models of discrimination: taste-based discrimination (Becker 1957) and statistical discrimination (Phelps 1972; Arrow 1973). Taste-based discrimination occurs when individuals act on personal biases or preferences against certain groups, leading to unequal treatment even when it is economically inefficient. This form of discrimination reflects a "cost" or disutility to engaging with individuals from the disfavored group. Statistical discrimination, by contrast, arises when decision-makers use group characteristics as proxies for individual attributes due to imperfect information. While market competition can, in principle, discipline taste-based discrimination, (accurate) statistical discrimination can persist even under perfect competition.

Recent work in economics offers extensions and refinements beyond the canonical taste-based and statistical models of discrimination. For example, Bohren et al. (2023) consider settings where decision-makers engage in inaccurate statistical discrimination—relying on group characteristics to infer individual traits, as in traditional statistical discrimination, but based on incorrect or biased beliefs about group averages. Another strand of research examines settings where the lack of individual-level data on quality gives individuals "moral wiggle room" to engage in taste-based discrimination that would be harder to sustain with more information. Other work emphasizes behavioral elements such as attention and stereotypes (Bordalo et al. 2016), as well as systemic and feedback-driven forms of discrimination (Bohren et al. 2025).

Understanding the source of discrimination provides insight into strategies to mitigate discrimination. For example, in a setting where statistical discrimination is the primary driver of bias, providing higher-quality information could reduce bias by eliminating the need to make decisions based on group characteristics.[1] However, if taste-based discrimination is at play, this information intervention will not fully eliminate bias.

---

[1] Altonji and Pierret (2001) formalize this intuition by showing, in the context of labor markets, that as firms observe more information about workers' true quality, the relationship between worker pay and easily observable characteristics like race should attenuate.

It is difficult to fully disentangle the various mechanisms that could explain an observed instance of discrimination (Bohren et al. 2023), but there is evidence suggestive of underlying mechanisms. For example, Nunley et al. (2011) document that racial discrimination against sellers of products on eBay diminishes as prospective buyers observe more information about sellers' quality via eBay feedback scores—a pattern consistent with statistical discrimination and the importance of information. Similarly, Doleac and Stein (2013) show that racial discrimination in the market for iPods on Craigslist is larger in areas with higher degrees of racial isolation and property crime. They note that buyers in these areas may perceive the cost of purchasing from a Black seller to be higher (e.g., because it is more likely that the sellers are criminals), which would indicate that statistical discrimination is a plausible mechanism.

Ge et al. (2020), by contrast, document that Uber drivers in Boston, Massachusetts, are more likely to cancel rides for Black customers, even though there was no true difference in passenger quality (as measured by star ratings) by race among the research assistants involved in the experiment.

Other evidence shows that information provision can meaningfully reduce bias, but may not be easily explained by standard models. For example, Chan (2024) shows that discrimination against Black and Asian doctors in an online healthcare marketplace decreases substantially when ratings about doctor quality are made visible. The racial willingness-to-pay gap falls by approximately 90% with the provision of this information. Additionally, users are willing to pay *more* for a Black doctor with the lowest quality ratings than for a Black doctor with no ratings—a result that appears inconsistent with accurate statistical discrimination. These patterns are more consistent with models that incorporate belief inaccuracy or that allow users to engage in bias when plausible deniability remains.

Thus, the empirical evidence is consistent with a variety of mechanisms being at work, depending on the setting studied. While models of statistical and taste-based discrimination are valuable, the evidence also suggests the need for an expanded set of models to understand the sources of bias and develop strategies to reduce bias.

3. Mitigating Discrimination and Promoting Equity

In response to growing evidence of discrimination on online platforms and in economic settings more broadly, researchers and companies have explored a range of strategies to reduce bias and promote equity. In this section, we review several broad categories of intervention: revealing or concealing demographic information, providing additional transaction-relevant information, modifying the platform's choice architecture, altering algorithmic decision rules, and introducing public policy responses. While the boundaries between these approaches are not always clear-cut, this classification offers a framework for understanding the range of potential responses and their underlying logic.

3.1 Revealing vs. Concealing Demographics

One key decision platforms face is whether to reveal or conceal users' demographic information. On the one hand, making traits like race easily visible can facilitate discrimination, as seen in many of the examples in the previous section. In response, some platforms have sought to reduce bias by obscuring demographic cues. For instance, Airbnb implemented a change that hid guests' profile pictures prior to booking, limiting the information that hosts could use to discriminate.

On the other hand, revealing demographic information can enable support from users who wish to engage in pro-equity behavior. A growing number of platforms—including Yelp, DoorDash, Spotify, Etsy, Target, Walmart, and Wayfair—have introduced programs that highlight businesses or creators from underrepresented groups. These include searchable tags (e.g., "Black-owned"), curated content hubs, product badges, and promotional placement. While the implementations differ, the shared goal is to help users discover and support minority-owned businesses and creators. Research suggests that such features can influence behavior, particularly in settings where demand for supporting marginalized groups is high.

Revealing demographic information can either enable discrimination or help channel support to minority users, depending on who uses that information and how. One factor in determining the effect of such features is selection: when platforms give users discretion over whether to engage with identity-relevant information or tools, the outcomes may reflect the preferences of those who opt in.

To illustrate this logic, consider a stylized application. Suppose that, in an effort to reduce discrimination against Black guests, Airbnb allows hosts to decide whether to adopt instant booking—a feature that lets qualified guests book immediately without requiring manual approval. Let U_{wi} and U_{bi} be the utilities host i derives from hosting a White guest and a Black guest, respectively. Define $B_i := U_{wi} - U_{bi}$ as host i's bias level. Assume that $B_i \sim \mathcal{N}(\mu, \sigma^2)$ where $\mu > 0$.[2] The existence of anti-Black bias on average is consistent with the results of Edelman et al. (2017). Additionally, define $p_w \in (0,1)$ to be the probability that a guest is White.

Consider an optional "instant-book" feature that gives qualified guests the ability to book listings immediately, without requiring manual approval from the host first. This feature may be attractive to hosts because it eliminates the need for them to manually look through applications, but giving up discretion might also be costly insofar as hosts have preferences over the set of applicants.

[2] We use a Gaussian distribution here for simplicity, but the results hold for any symmetric distribution, as shown in the Appendix.

To formalize this intuition, let host i's expected utility be $p_w \cdot U_{wi} + (1 - p_w) \cdot U_{bi}$ if they use the feature. If the host does not use this feature, their expected utility is given by $\max(U_{wi}, U_{bi}) - c$, where c is the cost of having to manually sift through applications.

Host i will opt to use this feature if and only if

$$p_w \cdot U_{wi} + (1 - p_w) \cdot U_{bi} > \max(U_{wi}, U_{bi}) - c$$

$$\Leftrightarrow$$

$$\frac{-c}{p_w} < B_i < \frac{c}{1-p_w}.$$[3]

This means that hosts whose bias is not too extreme adopt the technology. Intuitively, for these hosts, the cost of manual search outweighs the small benefit they receive from being able to rent to their preferred group. To formally see how the hosts who adopt the technology differ from the entire population of hosts, note that

$$\mathbb{E}\left[B_i \,\middle|\, \frac{-c}{p_w} < B_i < \frac{c}{1 - p_w}\right] < \mathbb{E}[B_i]$$

if and only if

$$c(2p_w - 1)/2p_w(1 - p_w) < \mu.$$[4]

The left-hand side of this expression is monotonically increasing in p_w. Thus, the condition is more likely to hold if μ is larger or p_w is smaller. Intuitively, with a sufficiently biased population on average and a sufficiently high likelihood that any given guest will be Black, users who adopt the "instant-book" feature are less biased on average than those in the population of hosts.

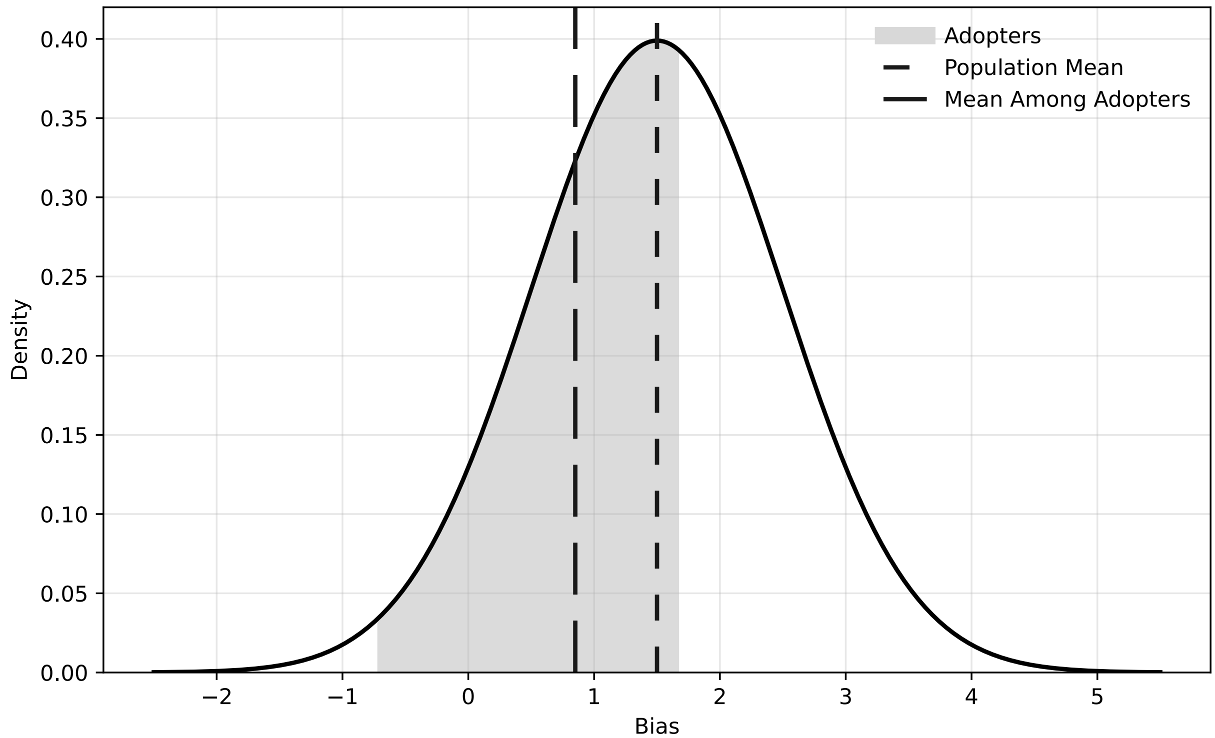
Figure 1 illustrates this visually for $\mu = 1.5$, $\sigma = 1$, $p_w = 0.7$, and $c = 0.5$. We observe that the shaded area, which represents the mass of hosts who adopt the instant-book feature, is concentrated toward the left (lower bias) side of the bias distribution and that the mean value of bias among these adopters is lower than the population mean.

---

[3] See the Appendix for the (short) derivation.

[4] See the Appendix for the derivation.

Figure 1: Bias in All Hosts vs. Instant-Book Adopters



Note: This figure presents a stylized model of host behavior under an "instant-book" feature that gives qualified guests the ability to book listings immediately. The figure plots the probability density function of the $\mathcal{N}(1.5,1)$ distribution to model host bias. In this example, we assume that the proportion of White guests p_w is 0.7 and that the cost of having to manually sift through applications c is 0.5. The shaded area shows the set of hosts who adopt the instant-book feature. The shorter dashed line shows the mean level of bias in the host population, while the longer dashed line shows the mean level of bias among the hosts who adopt the instant-book feature.

This result is relevant for a company like Airbnb, as it may be the high-bias hosts whose behavior the policy aims to alter. In a world where instant-book is made mandatory, by contrast, no host has the ability to discriminate. Allowing voluntary adoption influences outcomes not only through take-up but also through selection: adopters are likely to differ systematically in bias from non-adopters.

In other settings, though, selection may have different consequences. For example, in Aneja et al. (2025), the authors show that, in the context of restaurants on Yelp, allowing Black-owned restaurants to reveal the race of their owners via a Black-owned business badge increased demand for these businesses by channeling demand from users in low-bias areas. In that case, because information about businesses was not immediately available to consumers and must instead be learned with some search cost, only those sufficiently interested in supporting minority-owned businesses select into using the feature.

Likewise, in a variety of other settings outside of digital platforms, demographic information has been used by decision-makers interested in supporting discriminated-against groups. In the context of selection for the National Academy of Sciences and the American Academy of Arts and Sciences among psychologists, mathematicians, and economists, Card et al. (2023) found that, for the past twenty years, women's odds of selection have been higher than those for men, conditional on publications and citations. This is the reverse of the pattern observed before the 1990s, and the authors argue that it is suggestive of decision-makers aiming to help counteract bias faced by women in these fields. Similarly, Kirgios et al. (2022) show that local politicians are substantially more likely to respond to requests for help from Black and female senders when the senders explicitly reveal their demographic identity. In these contexts, as in the case of Yelp, suppressing demographic information would have shut down this channel for supporting minority groups, because the decision-makers using this information were inclined to support minority groups on average. This highlights the importance of platforms thinking carefully about the characteristics and preferences of the users who will observe demographic information if revealed.

3.2 Information Provision

Another key decision is what additional (non-demographic) information to provide. The optimal provision of such information depends in part on the source of discrimination on the platform. If discrimination is largely statistical in nature, providing information about user quality can be effective in reducing discrimination. For example, in the context of eBay, Nunley et al. (2011) show how for sellers with many feedback scores, the prices offered by prospective buyers did not vary depending on seller-buyer race combination. Even when mechanisms beyond accurate statistical discrimination are at play, providing reliable data on user quality can reduce biases. Recall that in the online marketplace studied by Chan (2024), buyers appear to hold substantially biased beliefs about the relationship between doctor race and quality. Even so, experimentally providing buyers access to information on doctor quality reduces the racial willingness-to-pay gap by 90%, highlighting the large effects that better information about user quality can yield.

Moreover, in the case where inaccurate beliefs are responsible for discrimination against a certain group, providing information about the true distributions of quality across groups can reduce discrimination.[5] For instance, Cullen et al. (2022) study hiring decisions on an online staffing platform in the U.S. and show that employers have inaccurate beliefs about the quality of workers with a criminal background: they believe workers with criminal backgrounds are more likely to be low-rated and less likely to be highly-rated than is actually the case. Providing information about the true share of high-performance ratings for these workers brought firms' beliefs more in line with reality and increased their willingness to hire by six percentage points (compared to a baseline of 39%). This effect on willingness to hire is approximately equivalent to the effect of subsidizing 40% of the

[5] Such an approach will generally not affect discrimination if beliefs are already accurate.

workers' wages, suggesting that the provision of group-level data may be an effective strategy to combat inaccurate statistical discrimination on platforms when providing individual-level data is infeasible or undesirable.

3.3 Choice Architecture

Holding the information provided constant, platforms have considerable flexibility in how they provide that information to users. A large literature in behavioral economics documents how choice architecture can affect consumer behavior (Leonard 2008). Platforms can leverage these insights to help achieve their goals of reducing bias and promoting equity.

Changes to the choice environment can help mitigate bias on digital platforms. For instance, Knittel et al. (2024), consistent with Ge et al. (2020), document that Uber drivers are more likely to cancel rides for Black passengers. They then implement an experiment modifying the design of the page drivers see when deciding whether or not to accept a ride. The information presented is the same across the treatment and control arms, but the size and brightness of the passenger's star rating were increased in the treatment, increasing its salience (Bordalo et al. 2012). The authors find that this treatment reduced the racial cancellation gap by over 70%.

As a second example, consider the case of Airbnb. One design feature Airbnb has experimented with is whether to make instant booking the default option or not. Making instant-book the default option has the potential to increase the overall usage of the feature, helping to counteract the selection into using the feature by less biased users that might take place otherwise, as described in Subsection 3.1. This default approach may be appealing to explore for platforms looking to reduce discrimination while not restricting hosts' choice set. Similarly, Airbnb has tested changing the default setting as to whether guests see only instant-book properties or all properties in a market and has experimented with removing host pictures from the main search results page. Neither change alters the information or choices available to guests but may nudge behavior toward a less-biased outcome. Of course, individual tweaks to the choice environment are unlikely to always produce large effects. Nudges should thus not be viewed as a panacea but rather as one tool in platforms' toolkit for mitigating bias and promoting fairness.

3.4 Other Design Choices

Platforms make many other design choices that do not fall neatly into any of the three categories listed above. Some such choices may seem at first glance to be unrelated to questions around discrimination and equity but may in fact have a substantial bearing on these issues. Consider the case of Hired.com studied by Roussille (2024). When a user creates a profile on Hired.com, they enter the base salary they aim to achieve. Prior to mid-2018, the box where one enters that information was blank. In mid-2018, the platform made a change so that the box would be pre-populated with the median salary for users with similar locations, job titles, and experience. This intervention was not designed with gender equity in mind, yet Roussille (2024) shows that it reduced the gender gap in asking wages and final offers to zero. This finding is consistent with literature documenting the gender gap in salary expectations (e.g., Reuben et al. 2017; Kiessling et al. 2024). In settings like this one, basic economic theory and relevant empirical literature can help platforms identify design choices that may affect bias and equity (in intended or unintended ways).

3.5 Algorithm Design

The previous four categories focused on the information provided and the design of the platform. Another aspect of the choice environment concerns the algorithms that shape users' experiences—for example, algorithms that determine what content users see and when. While a full discussion of algorithmic bias is beyond the scope of this article, we provide a few examples of where thinking about algorithm design may help promote fairness.[6]

To see the role that algorithms can play in shaping outcomes, consider Airbnb's decision about whether and how much to deprioritize listings of hosts who cancel on guests after accepting. If one thought that, as documented by Edelman et al. (2017) and others, Airbnb hosts would be biased against African Americans on average, a simple model of host decision-making would suggest that increasing the cost of canceling would reduce bias by increasing the penalty associated with discriminatory behavior. Additionally, guests may be less likely to book from biased hosts because their listings would be located less prominently. In this scenario, as in others, there may be reasons to implement or not implement a given algorithmic change, but knowing how such a change will differentially impact users is an input for platforms' decision-making processes.

Consider also the example of content delivery. Sweeney (2013) showed that individuals with stereotypically Black names were shown Google AdSense ads related to arrest records substantially more often than individuals with stereotypically White names. Whenever algorithms use user characteristics to determine what content to deliver, there is a potential for bias—namely, for otherwise identical users to be shown different content because of attributes such as gender or race. In some cases, these disparities may be intentional, for instance, when female users are disproportionately shown ads for women's clothing. In other cases, they may not be, underscoring the importance of ensuring that algorithms are not biased in harmful ways.

While algorithmic adjustments can help mitigate bias, doing so effectively requires careful attention to context. Even seemingly simple interventions—such as forcing equity in outcomes—can prove ineffective or even counterproductive if they fail to account for underlying objectives, data limitations, or unintended consequences. A recent example is

[6] See Dwork et al. (2011), Hardt et al. (2016), Kleinberg et al. (2016), Rambachan et al. (2020), Kasy (2024) for more detailed discussions about algorithmic fairness by computer scientists and economists.

Google's generative AI, Gemini, which produced historically inaccurate depictions (e.g., Black Nazis) in response to prompts intended to promote diversity. This case underscores the nuance involved in mitigating bias through algorithm design and the importance of evaluating interventions within their specific context.

3.6 Policy Responses

The previous subsections focused on tools that individual platforms can deploy to mitigate bias. Some of these (e.g., richer rating systems or redesigned choice architecture) can also improve efficiency and are therefore privately attractive. But platforms may not always be incentivized to combat discrimination. In such cases, government intervention has also been relevant. For example, in 2022, the U.S. Department of Justice settled with Meta over its housing-ad delivery algorithm, which allegedly used user characteristics protected under the Fair Housing Act. Likewise, regulators have begun to set ex ante rules for algorithmic hiring tools, most prominently with New York City's Local Law 144, which requires independent bias audits and public disclosure. Regulation in these instances has the potential to coordinate across platforms, for instance, by defining discrimination and by standardizing audit, disclosure, and record-keeping requirements.

Beyond disciplining platforms themselves, government regulators can apply and enforce existing civil rights laws on platform settings. Complaints can trigger agency investigations, data subpoenas, and audits. For example, in 2017, the California Department of Fair Employment and Housing announced an agreement with Airbnb that allowed the government to test individual hosts for racial bias via audit studies. Likewise, the National Federation of the Blind announced a 2017 settlement requiring Uber to deactivate drivers who refused riders with service animals and to compensate affected passengers. These actions illustrate how public enforcement, combined with platform data and enforcement tools, can directly target discriminatory behavior by users.

4. Measurement of Race

For platforms seeking to identify and address racial discrimination—and for researchers studying it—a key challenge is that users' race is often not explicitly observed. As a result, platforms and researchers rely on various methods to measure, infer, or proxy race from available data. This section briefly reviews several approaches to doing so.

4.1 Measurement Tools

One approach platforms can take is to ask users to optionally self-report race. This approach most directly captures self-identified race. However, this accuracy can come at the cost of scaling challenges: in practice, online surveys often have low completion rates, and thus many users may not fill out this information. For researchers, conducting surveys on platforms can also be challenging or infeasible.

A potentially more scalable way of inferring user race is by using proxies like users' names or profile pictures. A host of algorithms exist for inferring race from these sources. For example, the DeepFace software can be used to predict the race of a person in an image with just a few lines of code.[7] DeepFace outputs a vector of probabilities over possible races (Asian, White, Middle Eastern, Indian, Latino/Hispanic, and Black). Depending on the intended use, researchers can then choose whether to use these probabilities directly; to always use the race with the highest predicted probability, irrespective of its value; or to only use the algorithmic output when the maximum predicted probability exceeds some threshold (e.g., 80%). Algorithms like DeepFace have been deployed in academic research measuring racial bias on digital platforms. For example, Aneja et al. (2025) use DeepFace to measure the race of reviewers on Yelp to test whether the Black-owned business feature altered the racial composition of reviewers. They collect over one million photos from Yelp reviewers' profiles, use the DeepFace algorithm to predict reviewers' race, and use the model's output for downstream analysis.

Another approach employed in the literature is race imputation from names. Such algorithms are built using data sources, like the Census and voting records, attempting to document a ground-truth empirical relationship between names and (self-identified) race. A wide variety of these algorithms are easily accessible to researchers; they differ in the data used to build them and in the methods they employ. For example, in the programming language R alone, the package `wru` applies a Bayesian approach to data on surnames and geolocation from the Census; `rethnicity` uses a bidirectional LSTM on voting data from Florida; and `predictrace` uses the empirically observed proportions in Census data for last names and data from a sample of mortgage applications (Tzioumis 2018) for first names. As with the algorithms predicting race from images, these algorithms also typically output a vector of probabilities over possible races.

4.2 Considerations for Platforms

There are several considerations for platforms deciding which approach to measuring race to use.

Coverage. As discussed above, platforms and researchers may only be able to obtain self-reported race data for a small (self-selected) subset of users, rendering this approach more challenging for applications requiring race data for a large set (or the entire universe) of users. Depending on the nature of the platform, user names and/or photos may be observable for a larger set of users.

Accuracy. Different approaches will yield different accuracies in terms of how well measured race aligns with the ground truth. Insofar as self-identified race is the object of interest, then self-report data will by definition be perfectly accurate. Once one begins imputing race based on proxies like names or images, however, the possibility of error

[7] Information on the Python package for DeepFace can be found at the following link: https://github.com/serengil/deepface.

arises. While precise accuracy rates depend on the sample used and the algorithm deployed, there is some evidence comparing different approaches. For instance, Greenwald et al. (2024) compare the popular Bayesian Improved Surname Geocoding (BISG) approach to an image classifier in a dataset of applicants to the Paycheck Protection Program (PPP).[8] They find that race predictions from the image-based approach are substantially more correlated with true race than are predictions from the BISG approach ($r = 0.87$ vs. $r = 0.52$).

This kind of prediction error can be especially problematic if it is not random. In the context of an online platform for small business loans, Greenwald et al. (2024) show that non-Black applicants with lower levels of education are more likely to generate false positives, while Black applicants with higher levels of education are more likely to generate false negatives. The authors show that the non-randomness of these prediction errors means that an analysis of the racial gap in approval rates that uses race imputed from applicant names substantially understates the true gap.

Channel(s) of Discrimination. Another factor for platforms and researchers to consider is the channel through which discrimination might occur. Suppose, for example, that one is interested in studying racial discrimination against passengers on a ride-sharing app like Uber or Lyft, and one thinks that drivers may discriminate based on passenger name. In this setting, data on names may still be relevant, even if one has access to ground truth data on self-reported race. To see this, consider an imputation algorithm that mistakenly predicts that a subset of Black passengers are White because their names are not distinctively Black. Since drivers don't have access to passengers' true race, one would not expect them to discriminate against this subset of Black passengers, and so testing for racial discrimination on this subpopulation of passengers is unlikely to be illuminating. In this setting, one may be directly interested in discrimination against users with predictably Black names, so using race as inferred by names can be useful, especially insofar as drivers' assessments are similar to those of the algorithm.

5. Conclusion

This chapter has laid out several key points pertaining to equity on platforms. First, prejudice remains pervasive. We saw examples of bias against racial minorities in a wide variety of settings, ranging from Uber passenger pickups (Ge et al. 2020; Knittel et al. 2024) to doctor selection (Chan 2024).

Second, the canonical models of taste-based and statistical discrimination cannot always fully explain the discrimination observed in the data. For example, in the doctor selection

---

[8] The authors get applicant photos from LinkedIn. The authors build the image classifier used in the paper by using DeepFace to create embeddings (i.e., vector representations) of faces in a dataset of firm founders and then training a random forest classifier to predict race given the image embeddings.

setting studied by Chan (2024), the willingness-to-pay penalty customers impose on Black doctors is several times larger than the penalty imposed on the lowest-rated doctors. Patterns like this highlight the value of extending models to account for alternative explanations for the observed discrimination, such as inaccurate statistical discrimination (Bohren et al. 2023), where decision-makers hold biased beliefs, or a theory that posits that the absence of information on quality can provide "moral wiggle room" for taste-based discrimination. Carefully examining the source(s) of discrimination is important not only to decide on the appropriate remedy but also because doing so can allow existing theories to be tested in the field and give rise to new theoretical insights.

Third, platforms have a variety of tools at their disposal to address bias. One tool is the ability to decide what user information to make visible. Providing information about user quality (e.g., ratings of a doctor by past patients or of an Uber passenger by past drivers) may have the potential to reduce statistical discrimination (accurate or not) but may also have different—or smaller—effects if users are willing to pay to discriminate against minority groups. Providing information about user race can enable discrimination by revealing the key attribute on which discrimination occurs, yet in some settings (e.g., Aneja et al. 2025), doing so can channel demand from customers inclined to support minority groups. Thinking carefully about when and to whom to reveal race can shape the efficacy of such decisions. For example, if a platform has reliable signals of who is likely to support minority users, then it might be able to provide relevant information to these users. Allowing users to self-select into observing race may enable platforms to attract supportive customers. Yet as the stylized model provided in this chapter demonstrates, the effect can go the other way if those who opt out of observing information on race tend to be less biased. In addition to changing the information environment, platforms can also alter how that information is presented. For example, making information about user quality more salient may increase its effectiveness at reducing statistical discrimination (Knittel et al. 2024).

References

Acquisti, Alessandro, and Christina Fong. 2020. "An Experiment in Hiring Discrimination via Online Social Networks." *Management Science* 66 (3): 1005–24. https://doi.org/10.1287/mnsc.2018.3269.

Altonji, J. G., and C. R. Pierret. 2001. "Employer Learning and Statistical Discrimination." *The Quarterly Journal of Economics* 116 (1): 313–50. https://doi.org/10.1162/003355301556329.

Ameri, Mason, Sean Edmund Rogers, Lisa Schur, and Douglas Kruse. 2020. "No Room at the Inn? Disability Access in the New Sharing Economy." *Academy of Management Discoveries* 6 (2): 176–205. https://doi.org/10.5465/amd.2018.0054.

Aneja, Abhay, Michael Luca, and Oren Reshef. 2025. "The Benefits of Revealing Race: Evidence from Minority-Owned Local Businesses." *American Economic Review*, ahead of print. https://doi.org/10.3386/w30932.

Arrow, Kenneth J. 1973. "The Theory of Discrimination." In *Discrimination in Labor Markets*. Princeton University Press. https://doi.org/10.1515/9781400867066-003.

Ayres, Ian, Mahzarin Banaji, and Christine Jolls. 2015. "Race Effects on eBay." *The RAND Journal of Economics* 46 (4): 891–917. http://www.jstor.org/stable/43895621.

Bartlett, Robert, Adair Morse, Richard Stanton, and Nancy Wallace. 2022. "Consumer-Lending Discrimination in the FinTech Era." *Journal of Financial Economics* 143 (1): 30–56. https://doi.org/10.1016/j.jfineco.2021.05.047.

Becker, Gary S. 1957. *The Economics of Discrimination*. University of Chicago Press.

Bertrand, Marianne, and Sendhil 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. https://doi.org/10.1257/0002828042002561.

Bohren, J Aislinn, Peter Hull, and Alex Imas. 2025. "Systemic Discrimination: Theory and Measurement." *The Quarterly Journal of Economics* 140 (3): 1743–99. https://doi.org/10.1093/qje/qjaf022.

Bohren, J. Aislinn, Kareem Haggag, Alex Imas, and Devin G. Pope. 2023. "Inaccurate Statistical Discrimination: An Identification Problem." *Review of Economics and Statistics*, September, 1–45. https://doi.org/10.1162/rest_a_01367.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Stereotypes*." *The Quarterly Journal of Economics* 131 (4): 1753–94. https://doi.org/10.1093/qje/qjw029.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2012. "Salience Theory of Choice Under Risk." *The Quarterly Journal of Economics* 127 (3): 1243–85. https://doi.org/10.1093/qje/qjs018.

Broockman, David E., and Evan J. Soltas. 2020. "A Natural Experiment on Discrimination in Elections." *Journal of Public Economics* 188 (August): 104201. https://doi.org/10.1016/j.jpubeco.2020.104201.

Butler, Alexander W, Erik J Mayer, and James P Weston. 2022. "Racial Disparities in the Auto Loan Market." *The Review of Financial Studies* 36 (1): 1–41. https://doi.org/10.1093/rfs/hhac029.

Card, David, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri. 2023. "Gender Gaps at the Academies." *Proceedings of the National Academy of Sciences* 120 (4). https://doi.org/10.1073/pnas.2212421120.

Chan, Alex. 2024. "Discrimination Against Doctors: A Field Experiment." Harvard University.

Cullen, Zoë, Will Dobbie, and Mitchell Hoffman. 2022. "Increasing the Demand for Workers with a Criminal Record." *The Quarterly Journal of Economics* 138 (1): 103–50. https://doi.org/10.1093/qje/qjac029.

Doleac, Jennifer L., and Luke C. D. Stein. 2013. "The Visible Hand: Race and Online Market Outcomes." *The Economic Journal* 123 (572): F469–92. https://doi.org/10.1111/ecoj.12082.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. *Fairness Through Awareness*. https://arxiv.org/abs/1104.3913.

Edelman, Benjamin, Michael Luca, and Dan Svirsky. 2017. "Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment." *American Economic Journal: Applied Economics* 9 (2): 1–22. https://doi.org/10.1257/app.20160213.

Evsyukova, Yulia, Felix Rusche, and Wladislaw Mill. 2024. "Linkedout? A Field Experiment on Discrimination in Job Network Formation." *The Quarterly Journal of Economics*, ahead of print, October. https://doi.org/10.1093/qje/qjae035.

Ge, Yanbo, Christopher R. Knittel, Don MacKenzie, and Stephen Zoepf. 2020. "Racial Discrimination in Transportation Network Companies." *Journal of Public Economics* 190 (October): 104205. https://doi.org/10.1016/j.jpubeco.2020.104205.

Greenwald, Daniel L., Sabrina T. Howell, Cangyuan Li, and Emmanuel Yimfor. 2024. "Regulatory Arbitrage or Random Errors? Implications of Race Prediction Algorithms in Fair

Lending Analysis." *Journal of Financial Economics* 157 (July): 103857. https://doi.org/10.1016/j.jfineco.2024.103857.

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. *Equality of Opportunity in Supervised Learning*. https://arxiv.org/abs/1610.02413.

Kasy, Maximilian. 2024. "Algorithmic Bias and Racial Inequality: A Critical Review." *Oxford Review of Economic Policy* 40 (3): 530–46. https://doi.org/10.1093/oxrep/grae031.

Kiessling, Lukas, Pia Pinger, Philipp Seegers, and Jan Bergerhoff. 2024. "Gender Differences in Wage Expectations and Negotiation." *Labour Economics* 87 (April): 102505. https://doi.org/10.1016/j.labeco.2024.102505.

Kirgios, Erika L., Aneesh Rai, Edward H. Chang, and Katherine L. Milkman. 2022. "When Seeking Help, Women and Racial/Ethnic Minorities Benefit from Explicitly Stating Their Identity." *Nature Human Behaviour* 6 (3): 383–91. https://doi.org/10.1038/s41562-021-01253-y.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. *Inherent Trade-Offs in the Fair Determination of Risk Scores*. https://arxiv.org/abs/1609.05807.

Kline, Patrick, Evan K Rose, and Christopher R Walters. 2022. "Systemic Discrimination Among Large u.s. Employers." *The Quarterly Journal of Economics* 137 (4): 1963–2036. https://doi.org/10.1093/qje/qjac024.

Knittel, Christopher R., Donald MacKenzie, Michiko Namazu, Bora Ozaltun, Dan Svirsky, and Stephen Zoepf. 2024. *Using Information to Curb Racial Discrimination*. Working Paper No. 33118. National Bureau of Economic Research. http://www.nber.org/papers/w33118.

Laouénan, Morgane, and Roland Rathelot. 2022. "Can Information Reduce Ethnic Discrimination? Evidence from Airbnb." *American Economic Journal: Applied Economics* 14 (1): 107–32. https://doi.org/10.1257/app.20190188.

Leonard, Thomas C. 2008. "Richard h. Thaler, Cass r. Sunstein, Nudge: Improving Decisions about Health, Wealth, and Happiness." *Constitutional Political Economy* 19 (4): 356–60. https://doi.org/10.1007/s10602-008-9056-2.

Luca, Michael, Elizaveta Pronkina, and Michelangelo Rossi. 2024. "The Evolution of Discrimination in Online Markets: How the Rise in Anti-Asian Bias Affected Airbnb During the Pandemic." *Marketing Science*, ahead of print, February. https://doi.org/10.1287/mksc.2023.0112.

Munnell, Alicia H, Geoffrey M. B. Tootell, Lynn E. Browne, and James McEneaney. 1996. "Mortgage Lending in Boston: Interpreting HMDA Data." *American Economic Review* 86 (1): 25–53.

Nunley, John M., Mark F. Owens, and R. Stephen Howard. 2011. "The Effects of Information and Competition on Racial Discrimination: Evidence from a Field Experiment." *Journal of*

Economic Behavior & Organization 80 (3): 670–79. https://doi.org/10.1016/j.jebo.2011.06.028.

Pager, Devah. 2003. “The Mark of a Criminal Record.” *American Journal of Sociology* 108 (5): 937–75. https://doi.org/10.1086/374403.

Phelps, Edmund S. 1972. “The Statistical Theory of Racism and Sexism.” *The American Economic Review* 62 (4): 659–61.

Quillian, Lincoln, Devah Pager, Ole Hexel, and Arnfinn H. Midtbøen. 2017. “Meta-Analysis of Field Experiments Shows No Change in Racial Discrimination in Hiring over Time.” *Proceedings of the National Academy of Sciences* 114 (41): 10870–75. https://doi.org/10.1073/pnas.1706255114.

Rambachan, Ashesh, Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. 2020. “An Economic Perspective on Algorithmic Fairness.” *AEA Papers and Proceedings* 110 (May): 91–95. https://doi.org/10.1257/pandp.20201036.

Reuben, Ernesto, Matthew Wiswall, and Basit Zafar. 2017. “Preferences and Biases in Educational Choices and Labour Market Expectations: Shrinking the Black Box of Gender.” *The Economic Journal* 127 (604): 2153–86. https://doi.org/10.1111/ecoj.12350.

Roussille, Nina. 2024. “The Role of the Ask Gap in Gender Pay Inequality.” *The Quarterly Journal of Economics* 139 (3): 1557–610. https://doi.org/10.1093/qje/qjae004.

Stephens-Davidowitz, Seth. 2014. “The Cost of Racial Animus on a Black Candidate: Evidence Using Google Search Data.” *Journal of Public Economics* 118 (October): 26–40. https://doi.org/10.1016/j.jpubeco.2014.04.010.

Sweeney, Latanya. 2013. “Discrimination in Online Ad Delivery.” *CoRR* abs/1301.6822. http://arxiv.org/abs/1301.6822.

Tzioumis, Konstantinos. 2018. “Demographic Aspects of First Names.” *Scientific Data* 5 (1). https://doi.org/10.1038/sdata.2018.25.

Appendix

Proof of Results from Subsection 3.1

Adopting Condition. Let U_{wi} and U_{bi} denote the utilities from hosting a White or Black guest, respectively, and define

$$B_i \ := \ U_{wi} \ - \ U_{bi}.$$

A host adopts instant-book if and only if

$$p_w\, U_{wi} \ + \ (1-p_w)\, U_{bi} \ > \ \max(U_{wi},\, U_{bi}) \ - \ c,$$

where $p_w \in (0,1)$ is the probability of a White guest and $c > 0$ is the cost of manually reviewing applications. We can express this condition in terms of B_i by observing that

$$\begin{aligned} p_w\, U_{wi} \ + \ (1-p_w)\, U_{bi} &= U_{bi} + p_w(U_{wi} - U_{bi}) \\ &= U_{bi} + p_w B_i \end{aligned}$$

and

$$\begin{aligned} \max(U_{wi}, U_{bi}) &= \max(U_{bi} + B_i,\, U_{bi}) \\ &= U_{bi} + \max(B_i, 0). \end{aligned}$$

Hence the original inequality

$$p_w\, U_{wi} + (1-p_w)\, U_{bi} > \max(U_{wi}, U_{bi}) - c$$

becomes

$$U_{bi} + p_w B_i > U_{bi} + \max(B_i, 0) - c$$

which is equivalent to

$$p_w\, B_i + c > \max(B_i, 0).$$

If $B_i > 0$, this condition simplifies to

$$p_w B_i + c > B_i \Leftrightarrow B_i < \frac{c}{1-p_w}.$$

If $B_i < 0$, this condition simplifies to

$$p_w B_i + c > 0 \Leftrightarrow B_i > \frac{-c}{p_w}.$$

If $B_i = 0$, this condition simplifies to

$$c > 0,$$

which is true by construction. Thus host i chooses to adopt the instant-book technology if and only if

$$\frac{-c}{p_w} < B_i < \frac{c}{1-p_w}.$$

Bias of Instant-Book Adopters vs. Average Host. Let B_i be a random variable with mean μ, and assume B_i is symmetric about μ. Set

$$a \;=\; -\frac{c}{p_w}, \quad b \;=\; \frac{c}{1-p_w}, \quad m \;=\; \frac{a+b}{2} \;=\; \frac{c(2p_w-1)}{2p_w(1-p_w)}$$

and define $X := B_i - \mu$. By assumption, X is symmetric about 0. Truncating B_i to (a, b) is equivalent to truncating X to $(a-\mu,\, b-\mu)$. Hence,

$$\mathbb{E}[B_i \mid a < B_i < b] \;=\; \mu \;+\; \mathbb{E}[X \mid a-\mu < X < b-\mu].$$

Set $\alpha = a - \mu$ and $\beta = b - \mu$. Then the midpoint m corresponds to

$$m \;-\; \mu \;=\; \frac{\alpha+\beta}{2}.$$

By symmetry, we know

$$\mathbb{E}[X \mid \alpha < X < \beta] > 0 \quad \Leftrightarrow \quad \frac{\alpha+\beta}{2} > 0.$$

Thus,

$$\mathbb{E}[\, B_i \mid a < B_i < b] \;=\; \mu + \mathbb{E}[X \mid \alpha < X < \beta] \;<\; \mu \quad \Leftrightarrow \quad m < \mu.$$