Systemic Discrimination: Theory and Measurement[*]

J. Aislinn Bohren[†] Peter Hull[‡] Alex Imas[§]

October 12, 2024

[*] We thank Jason Abaluck, Amanda Agan, David Arnold, Marianne Bertrand, Peter Blair, Leonardo Bursztyn, Hanming Fang, Damon Jones, Diag Davenport, Stefano DellaVigna, Will Dobbie, Ed Glaeser, Sam Hirshman, Larry Katz, Erzo Luttmer, Matthew Knepper, Matt Lowe, Muriel Niederle, Matthew Notowidigdo, Anna Gifty Opoku-Agyeman, Jane Risen, Evan Rose, Heather Sarsons, Andrei Shleifer, Fabio Tufano, Damián Vergara, Basit Zafar, and seminar participants and conference participants at various institutions for helpful comments. Cuimin Ba, Joshua Hascher, Christy Kang, Matthew Murphy and Yier Ling provided expert research assistance. Hull gratefully acknowledges funding through the National Science Foundation (Award #2119849).

[†] Department of Economics, University of Pennslyvania: abohren@sas.upenn.edu

[‡] Department of Economics, Brown University: peter_hull@brown.edu

[§] Booth School of Business, University of Chicago: alex.imas@chicagobooth.edu

1 Introduction

Disparities by race, gender, and other protected characteristics are widely documented, raising concerns of discrimination. In economics, such concerns are usually probed with a rich set of tools for modeling and measuring *direct* discrimination: how protected characteristics affect individual actions, holding fixed all other relevant factors. An economist might, for example, measure direct discrimination by estimating the causal effect of a hiring manager's perceptions of a job applicant's race, holding fixed the applicant's work experience and education. She might interpret any such race effects through canonical models of taste-based or statistical discrimination, or other theories of direct discrimination.

There is, however, a growing recognition that focusing on direct discrimination can yield an incomplete understanding of how societal inequities can arise, persist, and compound. Sociologists and legal scholars have long emphasized the importance of systems-based analyses, which study discrimination as the cumulative outcome of interactions across different periods and domains (Pincus 1996; Powell 2007; De Plevitz 2007; Small and Pager 2020). Decades of research in labor economics has similarly noted how "pre-market" discrimination in education and housing systems might affect the employment opportunities of minorities even in the absence of direct discrimination (Cain 1986; Neal and Johnson 1996; Bertrand and Mullainathan 2004a).[1] More recently, computer scientists have shown how discrimination in algorithmic decisions can arise indirectly from biased data collection and training systems even when the algorithm is "blinded" to protected characteristics (Angwin, Larson, Mattu, and Kirchner 2016; Rambachan and Roth 2020). Yet despite these literatures, when compared with the robust toolkit for modeling and measuring direct discrimination, economists have more limited theoretical and empirical methods to study such indirect or *systemic* forms of discrimination (Small and Pager 2020).

This paper develops a common theoretical framework for studying direct and systemic discrimination and new tools for bringing this framework to economic data. We focus on a notion of systemic discrimination that captures how direct discrimination in other decisions leads to differences in relevant attributes for a given decision, which in turn generates disparities in outcomes. Our framework contributes to the sociology and legal literatures a precise mathematical language for analyzing such systemic discrimination and its drivers.[2]

[1] For example, Cain (1986) presents two statistical models for measuring discrimination: model (I) identifies discrimination as group-based differences in an outcome variable of interest, controlling for all relevant productivity characteristics; model (II) identifies discrimination as the unconditional difference in the outcome variable and implicitly attributes any differences in productivity characteristics to pre-market discrimination.

[2] As Small and Pager (2020) note, terms like "systemic" or "structural" discrimination—while broadly referring to the idea that "something other than individuals may discriminate"—are often imprecisely and inconsistently used across the social sciences. We provide formal definitions of direct, systemic, and total discrimination and discuss how these definitions map onto existing concepts in the literature.

To the labor economics literature on pre-market discrimination and computer science literature on algorithmic fairness, we contribute a general approach for formally modeling systemic discrimination in a wide range of settings. Importantly, adding structure to the notion of pre-market discrimination yields a novel empirical strategy—the *Iterated Audit*—that builds on existing experimental methods to measure both forms of discrimination within a pre-defined "system." We show how this approach can be used to quantify the impact of systemic discrimination on observed disparities and to help inform potential policy responses.

We start by developing the general framework for examining direct and systemic discrimination, including how the former can contribute to the latter through interactions across time and contemporaneous domains. We model a "system" as a network of interconnected nodes, each representing a decision (e.g. a hiring manager considering the individual for a job) that can affect an individual's relevant attributes (e.g., work experience or education) at other decision nodes. We show how disparities at a given decision node arise from three distinct channels: direct discrimination at that node (e.g., preferring to hire men over women with identical experience and education), systemic discrimination arising from direct discrimination at other nodes (e.g., direct discrimination by teachers leading to differential educational attainment for women vs. men and hence, hiring disparities), and differences in characteristics that arise outside of the system (e.g., gendered differences in innate physical capacity for work). Systemic discrimination can arise from either informational or technological sources, depending on whether direct discrimination at other nodes generates differences in the signaling technology (e.g., education signals the worker's productivity) or differences in payoff-relevant characteristics (e.g., education increases worker productivity itself). Aggregating disparities from the direct and systemic channels at a given node corresponds to what we refer to as *total* discrimination.

The framework highlights a key analytic choice that a researcher interested in quantifying systemic discrimination must make: which systemic forces to study. This amounts to a choice of which other decision nodes to include in the analysis, in addition to the "focal" node at which discrimination is being measured. For example, a researcher interested in studying systemic discrimination in entry-level hiring (the focal node) stemming from direct discrimination by reference letter writers would analyze decisions at both the entry-level hiring and reference letter nodes. This would allow her to measure how direct discrimination in reference letters translates to disparities in entry-level hiring. A different researcher interested in studying all informational sources of systemic discrimination—by, for example, conditioning on a worker's ex-post observable productivity at the focal node—would instead include all nodes that generate disparities in the worker's ability to signal her human capital. Both approaches contrast with a researcher only interested in direct discrimination in hiring, who would implicitly include only a single node—the focal node—in the analysis. Thus, by choosing different sets of nodes, the framework provides a unified structure that nests different notions of discrimination—breaking down the complex system that generates disparities into interpretable and measurable components. There may be one or several natural choices in any given setting, depending on the systemic forces of interest to the researcher, data availability, or the scope for policy interventions.[3]

We next show how this framework can be brought to data with the iterated audit (IA) approach. An IA analysis involves two key components: (*i*) a treatment component capturing direct discrimination at both the focal node and other included nodes, and (*ii*) an interaction component capturing how actions at the other included nodes impact treatment at the focal node. For example, consider a two-node setting in which individuals first apply for an internship and then an entry-level position (the focal node). An IA analysis involves measuring (*i*) direct discrimination in internship hiring for individuals with similar resumes, (*ii*) direct discrimination in entry-level hiring for individuals with similar resumes including internship experience, and (*iii*) how internship experience impacts entry-level hiring for individuals with otherwise similar resumes. The latter interaction component shapes how direct discrimination at other nodes drives systemic discrimination at the focal node.

Direct discrimination, and hence the IA treatment component, can be identified by conventional experimental designs (e.g., audit or correspondence studies). We develop two identification strategies for the interaction component, and hence, systemic discrimination. The *constructive* approach separately estimates the impact of each possible action at another node on treatment at the focal node; combining these estimates identifies the interaction component. The *experimental* approach directly measures systemic discrimination by simulating the distribution of focal-node signals, as impacted by group membership and the actions at other focal nodes, and measuring focal-node actions given the simulated signals. For example, the interaction component above could be constructed from estimates of how having internship experience affects entry-level hiring rates by race for individuals with otherwise similar resumes. Alternatively, one could simulate internship experience by race and measure subsequent entry-level hiring. The latter experimental approach may be preferred when attributes are high-dimensional or otherwise complex (e.g., text data), making estimation of the interaction component difficult. Both approaches improve over simpler alternatives—such as those that simply add conventional direct discrimination estimates across nodes without measuring the interaction component—particularly when actions at other nodes impact decisions at the focal node non-linearly or when decision-makers act to

[3] Most analyses of systemic discrimination in sociology can be understood as choosing a measure that includes all possible nodes. This also corresponds to model (II) in Cain (1986). As they note, however, such measures captures both obvious discrimination and what might be viewed as non-discriminatory differences in attributes (e.g., innate physical characteristics). This paper shows how more structured measures of systemic discrimination can distinguish between these two components.

undo any perceived direct discrimination at other nodes.

We illustrate these new tools in two field experiments. The first used the constructive IA approach to study how direct racial discrimination in entry-level job hiring can generate systemic discrimination in later hiring via disparities in applicant work experience. We studied a system with two nodes. Applicants apply for jobs without prior work experience in the first round of hiring (the first node) and either obtain a job or not. They then apply for a job in the second round of hiring (the second, focal node), with or without prior experience from the first round. Direct discrimination in first-round hiring can thus become embedded in work experience, leading to systemic discrimination in second-round hiring. We used a correspondence study to estimate direct discrimination in both rounds, then used additional data to construct the interaction component and estimate systemic discrimination.

Specifically, we built on the correspondence study methodology (Bertrand and Mullainathan 2004b; Kline, Rose, and Walters 2021) by generating job applications for a fictitious group of workers that vary in their level of experience and submitted them to online job vacancies at a set of national firms. We focused on the automotive firms which Kline et al. (2021) document as having the highest levels of direct discrimination in callback rates, and similarly randomize applicants' names to signal their race.[4] In the first round of hiring, we found sizable direct discrimination: among those with no previous work experience, applicants with distinctively white names were 13 percentage points (90%) more likely to receive a callback than applicants with distinctively Black names.

Estimating the interaction component, and hence systemic discrimination, requires (i) assessing the return to experience in the second round of hiring and (ii) how direct discrimination in the first round translates to race-based differences in experience. For the former, we sent out resumes that had one line of previous experience—at similar firms as the target job—which differed only in the name of the applicant. We found substantial returns to experience: overall, applicants with one line of previous experience were 10 percentage points (50%) more likely to receive a callback than those without. This suggests a meaningful role for systemic discrimination, as direct discrimination in the first round affects an important attribute (experience) for second-round hiring. How this direct discrimination translates depends on local market thickness—i.e., the number of jobs a first-round applicant can apply to—and the rate at which first-round callbacks convert to employment. To estimate market thickness, we scraped the number of job openings across the municipalities in our experiment. Finally, we estimated callback conversion rates by surveying a separate sample of hiring managers in the automotive industry.

[4] Kline et al. (2021) ran a correspondence study across a large set of US firms. The study used fictitious resumes that were the same apart from the name of the applicant, which was either distinctly Black or distinctly White. Importantly, each resume had at least some level of experience. The authors observed a significant 6 percentage-point call-back gap between Black and White resumes sent to automotive firms.

Combining these data, we found significant systemic discrimination in second-round hiring—comprising roughly half of the measured total discrimination. The other half was due to direct discrimination, which was lower in the second round due to lower direct discrimination against experienced applicants (4 percentage points compared to the 13 percentage points documented for inexperienced applicants). This implies that simply looking at the conventional direct measure would have underestimated total discrimination by 50%, highlighting the importance of these new tools for measuring the full extent of racial inequity. Our results also illustrate the utility of the IA method for assessing potential policy responses. The size of second-round systemic discrimination suggests that targeting first-round direct discrimination would significantly mitigate observed disparities in subsequent rounds. Moreover, the role of local market thickness in shaping systemic discrimination suggests scope for further policy targeting: systemic discrimination is lower in markets with low and high levels of thickness, implying that policy directed to areas with intermediate levels of thickness would be most effective for mitigating the compounding impacts of direct discrimination.

Our second experiment demonstrates the value of the experimental IA method in settings with high-dimensional or complex signals. Here we measured how direct gender discrimination in the language of recommendation letters can generate systemic discrimination in hiring. We again studied a system with two nodes. Applicants received recommendation letters based on their resumes (the first node), then apply for a job (the second, focal node) by submitting their resumes and recommendation letters. Prior work has found significant language differences in the letters of similarly qualified male and female applicants (e.g. Schmader, Whitehead, and Wysocki 2007), which are replicated in automated recommendation letters from large language models (LLMs) (Wan, Pu, Sun, Garimella, Chang, and Peng 2023). Direct discrimination in recommendation letter language thus becomes a part of the job-seekers' application materials. But it is not obvious how such language differences may contribute to subsequent hiring disparities: hiring managers may focus on "hard" information contained in the resume and ignore the differences in recommendation letters, or these differences may not lie in language that is most relevant for the hiring decision. Estimating these interactions is challenging because of the high-dimensionality of text data.

To quantify the systemic impact of recommendation language disparities on labor market outcomes, we randomized distinctively male or female names to a set of fictitious resumes and generated recommendation letters from them via standard LLM-based techniques. As in prior work, the resulting recommendation letters displayed marked gender-based differences in language. Following the experimental version of the IA method, we then generated three sets of "materials" (resumes and recommendation letters). The first two sets (A and B) were as in a standard correspondence study, with set A assigned distinctively male names, set B assigned distinctively female names, and the recommendation letters of both held fixed as the ones generated for male candidates. The third set C was assigned female names and recommendation letters generated for female candidates. We submitted these materials to a set of real-world hiring managers and elicited expected hiring probabilities and wages via an incentivized ratings design (Kessler, Low, and Sullivan 2019). A comparison of outcomes for set A vs. B thus identifies direct gender discrimination, holding fixed applicant materials, while a comparison of B vs. C identifies systemic discrimination from the direct discrimination in recommendation letters. Taken together, a comparison of A vs. C identifies total discrimination inclusive of systemic language disparities in the letters.

We find that essentially all of the gender discrimination in hiring rates and wages is driven by systemic disparities from recommendation letter language. Overall, applicants with distinctively male names and male recommendation letters were substantially more likely to be hired and were assigned a 21% higher wage than applicants with distinctively female names and female recommendation letters. Holding recommendation letters fixed, however, shrinks the hiring and wage disparities to insignificant levels.

These results speak to recent work on gender discrimination in labor markets which suggests a limited role for direct discrimination in explaining observed gender disparities in the labor market. For example, a recent meta-analysis of correspondence studies has found little support for discrimination in recent years when the non-group information of male and female candidates is held fixed (Schaerer, Du Plessis, Nguyen, Van Aert, Tiokhin, Lakens, Clemente, Pfeiffer, Dreber, Johannesson et al. 2023); Ceci and Williams (2011) present similar results in the case of academic jobs. Our results suggest systemic discrimination can arise from direct discrimination in signaling, for example, how male and female candidates are described to potential employers, can potentially explain some of the total disparities observed in the labor market that conventional direct discrimination measures would miss.

This paper builds on several related literatures in economics. The notion of pre-market discrimination in labor economics is captured as a form of systemic discrimination in our framework, and some models in this literature (e.g. Coate and Loury (1993); Cornell and Welch (1996)) microfound a particular system of nodes. Our theoretical framework nests these models, capturing both broader notions of systemic discrimination from outside of economics (e.g. Feagin (2013); Gynter (2003)) as well as more modern notions of indirect economic discrimination (e.g. Hurst, Rubinstein, and Shimizu 2021). Connecting these different literatures yields a general approach to measurement and a unified framework for considering appropriate policy responses.

Empirically, our approach relates to a concern from the economics discrimination literature of "bad controls"—i.e., conditioning on characteristics that are themselves affected by discrimination (e.g. Cain 1986; Altonji and Blank 1999). We show how specifying the forms of systemic discrimination of interest in an analysis, via the modeled nodes, leads naturally to the specification of "good" and "bad" controls in our IA method. Our IA approach further shows how discrimination measures with more and fewer controls can be reconciled as identifying different forms of direct and total discrimination, and how systemic discrimination can be inferred by their difference.

Finally, our experimental findings add to a small but growing literature estimating the impact of previous direct discrimination on subsequent disparities (Cook 2014; Williams, Logan, and Hardy 2021; Eli, Logan, and Miloucheva 2023; Derenoncourt, Kim, Kuhn, and Schularick 2022; Harrington and Shaffer 2023. A series of recent empirical papers also build directly on our framework to measure and classify direct, systemic, and total discrimination in various settings (Althoff and Reichardt 2022; Baron, Doyle Jr, Emanuel, Hull, and Ryan 2023; Zivin and Singer 2023; Lodermeier 2023; Gawai and Foltz 2023; Buchmann, Meyer, and Sullivan 2023; Conway, Mill, and Stein 2023). More broadly, we join a growing literature modeling and estimating the indirect impact of discrimination on important economic outcomes—including Darity (2005), Bohren, Imas, and Rosenberg (2019), Bohren, Haggag, Imas, and Pope (2022), Arnold, Dobbie, and Hull (2022), and Hurst et al. (2021).

2 Motivating Example

We begin with a simple theoretical example that illustrates the key features of the framework.[5] Consider a population of patients i seeing a physician in order to decide whether to get a colorectal cancer screening. The goal of such screenings is to detect early signs of cancer. Correspondingly, let $Y_i^* \in \{0, 1\}$ indicate the latent presence of cancer in patient i.

Before seeing the physician, each patient is first seen by a nurse practitioner (NP). The NP takes down the patients demographics (including self-reported race, G_i), checks their basic medical information (e.g. height, weight, blood pressure), and conducts a short medical history survey. The survey includes a variety of open-ended questions on the patient's experience with screening and cancer, and NPs have some discretion as to how they record a patient's answers. The physician receives a file from the NP with all the collected information, conducts their own short interview, then makes a screening recommendation. Let S_i denote all information available to the physician for patient i excluding the self-reported race, and let A_i denote her action (a screening recommendation).

A large literature has found substantial race-based disparities in screening decisions (e.g., Jerant, Fenton, and Franks (2008), Crawley, Ahn, and Winkleby (2008)); three economists are interested in studying the role of discrimination in explaining these disparities. Economist 1 follows standard practices in the field by designing and conducting a careful audit study. Specifically, she recruits a set of white and Black patients with comparable demographics and basic medical information, and randomizes them to pre-filled NP files and scripts for

[5] This example is inspired by Zink, Obermeyer, and Pierson (2023).

interacting with the physician. In this way, she ensures identification of the effect of race G_i on the action A_i conditional on the physician's non-race signals S_i. She finds that white patients are, on average, somewhat less likely to receive a screening recommendation than Black patients assigned to the same file and script. She concludes that there is some racial discrimination against white patients in this setting.

Economist 2 is interested in the same question, but ends up running a somewhat different audit study. Rather than randomizing NP files directly, she randomizes scripts for the recruited white and Black patients to interact with *both* the NP and physician. That is, while she ensures white and Black patients have the same screening and cancer history to report to the NP, she allows NPs to affect the recording of this information that is given to the physician through S_i. Strikingly, this disparity in the design yields a different conclusion than Economist 1's: white patients are, on average, slightly *more* likely to receive a screening recommendation than Black patients assigned to the same set of scripts. Thus she concludes there is some racial discrimination against Black patients in this setting. In unpacking this result, she finds that NPs tend to use more serious language in recording the history of white patients relative to Black patients randomized to the same family history.

Finally, Economist 3 examines the same question by running a different type of study. Randomly screening a representative set of white and Black patients after the physician makes a recommendation, she measures true rates of cancer incidence Y_i^*. This allows her to compute racial disparities in screening recommendations among patients with the same cancer status, without conditioning on any non-race signals. Curiously, she reaches a different conclusion than both Economist 1 and 2: white patients are *much* more likely to receive a screening recommendation than Black patients with the same underlying cancer status. In unpacking this result, Economist 3 finds that a key driver is the differential accuracy of available family history information by race: among patients with the same cancer status, Black patients are much less likely to know whether their parents or grandparents suffered from colorectal or related cancers due to more limited historical interactions with doctors.[6]

At first blush, this simple example presents a puzzle: which of the three researchers are correct on the nature and extent of discrimination in cancer screening recommendations? Economist 1 follows the norm in economics by conditioning on the information available to the decision-maker, S_i. Economist 2 finds that some of this signal is biased by a different stage of the recommendation system (i.e. the language disparity in the NP's notes); her measure of discrimination takes a starting point before the patient's current interaction with the healthcare system. By conditioning on an "objective" measure of qualification for screening—the underlying cancer risk Y_i^*—Economist 3 uncovers a further bias in the patient's recorded family history. How can such *systemic* biases be coherently studied alongside the *direct*

[6] See, e.g., Kupfer, McCaffrey, and Kim (2006).

discrimination that Economist 1 documents in physician decision-making?

In Section 3 we develop a general framework for reconciling these different analyses. The framework formalizes how the study of discrimination requires a researcher to take a stance on the notion of "qualification" for a given decision—what we call Y_i^0—which is the key factor that differed across the three researchers. In any given setting, there may be one or several natural choices for Y_i^0; by selecting different reference qualifications, a researcher can study different systemic forces alongside canonical sources of direct discrimination.

Two other points, which we return to in subsequent sections, are worth highlighting in this example. First, as shown in Section 4, studying such systemic forms of discrimination generally requires new empirical tools. While Economist 1 identified direct discrimination with a standard audit experiment, isolating the effects of the systemic biases found by Economists 2 and 3 is more challenging. We propose an alternative iterated audit design to identify these effects, and discuss how applying this design may require different (quasi-) experimental designs—particularly when the chosen qualification reference point is imperfectly observed.

Second, it may be difficult to address such systemic forms of discrimination with standard individual-level interventions. The physician in this example seems at least partly aware of the systemic bias in NP notes, given her "reverse" discrimination in recommending screenings at a lower rate for white patients than observably-similar Black patients. Yet she does not fully offset the bias, either because of imperfect awareness or because of her own psychological frictions or biases. Hence broader or system-wide policy responses may be called for in settings with significant systemic discrimination. By nesting different forms of discrimination in a single framework, our approach can be used to formulate and target such systems-wide policy responses and to study how they may impact other interconnected decisions.

3 Formalizing Systemic Discrimination

We develop a theoretical framework to put structure on how to measure systemic and total discrimination. Section 3.1 introduces the setting, Section 3.2 presents the definitions of discrimination, Section 3.3 shows how this framework nests common notions of discrimination in the literature, and Section 3.4 discusses key features of the framework and delineates two sources of systemic discrimination.

3.1 Framework

We develop our framework in a labor market context, in which we consider discrimination towards a worker being evaluated for a task n^*. Worker i has ex-ante unobservable *productivity* $Y_i^* \in \mathcal{Y}$ on the task, where $\mathcal{Y}^* \subset \mathbb{R}$ is the set of possible productivity levels and Y_i^* is a measure of the worker's capacity to complete the task (e.g., error rate, number of units produced). A manager observes the worker's *group* identity $G_i \in \{b, w\}$ and a vector of attributes $S_i^* \in \mathcal{S}^*$ (e.g., educational background, recommendation letters, etc.) which we refer to as the *signal*, where $\mathcal{S}^*$ is the set of possible values for each attribute. A worker's attributes and (potentially) group identity provide information about the worker's productivity Y_i^*. The manager then evaluates the worker by selecting an *action* $A_i^* \in \mathcal{A}^*$ (e.g., hiring decision, offered salary, performance rating), where $\mathcal{A}^* \subset \mathbb{R}$ is the set of possible actions. The manager's payoff depends on her action, the worker's productivity, and (potentially) group identity. She maximizes her expected payoff subject to her beliefs about the joint distribution of productivity, the signal, and group identity. Rather than explicitly modeling the manager's decision problem, we take a reduced-form approach by specifying the manager's decision rule $A^* : \mathcal{S}^* \times \{b, w\} \to \mathcal{A}^*$, which determines how the observed signal and group identity maps into an action choice. Given G_i and S_i^*, the manager selects action $A_i^* = A^*(G_i, S_i^*)$.[7]

We embed this employment evaluation in a broader economy—a *system*—to capture the idea that the worker's productivity and signal for task n^* may be affected by decisions in other domains and time periods (e.g., financial, criminal, past employment). A system consists of a set of nodes $\mathcal{N} \equiv \{1, ..., N\} \cup \{n^*\}$, where each node $n = 1, ..., N$ corresponds to a task similar in structure to the one described above. Specifically, at each node n, a worker is evaluated for a task for which he has productivity $Y_i^n \in \mathcal{Y}^n \subset \mathbb{R}$ (e.g., loan repayment probability, criminal activity propensity, past job performance). An evaluator (e.g., loan officer, judge, past manager) observes the worker's group identity G_i and a signal $S_i^n \in \mathcal{S}^n$, then selects action $A_i^n \in \mathcal{A}^n \subset \mathbb{R}$ (e.g., loan terms, criminal charges, hiring decision). Note that group identity G_i is not task-dependent; it is fixed for the worker across all nodes in the system. As above, Y^n, $\mathcal{S}^n$, and $\mathcal{A}^n$ describe the set of possible values of productivity, the signal, and action, respectively, and $A^n(G_i, S_i^n)$ denotes the evaluator's decision rule.

Evaluations at a given node can impact the productivity and signal at other nodes. For example, the productivity Y_i^* at node n^* may be a function of past productivity Y_i^n and on-the-job training $A^n(G_i, S_i^n)$ at node n, where $A^n(G_i, S_i^n)$ captures the allocation of such training. Similarly, the signal S_i^* at node n^* may contain a past performance evaluation $A^n(G_i, S_i^n)$ at node n, where $A^n(G_i, S_i^n)$ captures this evaluation. Given this system, we refer to node n^* as the *focal node* since it is the node at which discrimination is being studied.

3.2 Definitions

We define three forms of discrimination: direct, total and systemic. *Direct discrimination* captures group-based differences in action choices at a given node, holding fixed the signal. It occurs when the action rule prescribes different actions for group w and b workers with

[7] We abstract from interactions across workers and other realistic features of labor markets for simplicity. We assume productivity and the action are real numbers and the set of groups is binary to simplify notation; the analysis easily extends to more general spaces.

the same signal realization:

Definition 1 (Direct Discrimination)**.** *Direct discrimination occurs at node $n \in \mathcal{N}$ for signal $s \in \mathcal{S}^n$ if $A^n(w, s) \neq A^n(b, s)$.*[8]

Direct discrimination arises from the worker's group identity itself; it is a causal concept because it conditions on all observed non-group attributes. It can arise from the dependence of the evaluator's preferences, beliefs about productivity, or beliefs about the signal distribution on group identity.

Our definitions of systemic and total discrimination capture a broader notion of inequity that incorporates how direct discrimination in other decisions contribute to disparities in the present one. A key analytic choice that a researcher interested in quantifying such broader inequities must make is which systemic forces to study. The researcher may wish to hone in on a subset of nodes in order to isolate the impact of direct discrimination within this subset on disparities at the focal node—perhaps because detailed decision data is available for these nodes, policy-makers seek to intervene at these nodes, or simply because these are the systemic forces of interest to the researcher. This choice amounts to specifying which other nodes to include in the analysis. Let $\mathcal{N}^0 \subset \mathcal{N}$ denote this chosen set of nodes, which we refer to as the *subsystem*. A worker i enters the subsystem with *reference qualification* $Y_i^0 \in \mathcal{Y}^0 \subset \mathbb{R}$, which summarizes the components of productivity Y_i^* and signal S_i^* at the focal node that are generated outside of the subsystem. This controls for disparities that stem from nodes outside the subsystem, thereby serving as a reference point from which one can measure how systemic forces accumulate within the subsystem.

For example, a researcher interested in studying how direct discrimination in entry-level hiring contributes to disparities in promotion (the focal node n^*) would analyze decisions at both the entry-level and promotion nodes (i.e., the subsystem consists of these two nodes). This amounts to excluding other nodes at which the worker may have faced direct discrimination that impacted promotion (e.g., prior education). The reference qualification Y_i^0 corresponds to the components of the worker's productivity Y_i^* and signal S_i^* at the promotion node that arise from decisions other than entry-level hiring. For example, if entry-level hiring adds a line of experience to a worker's resume (the signal S_i^* includes $A^n(G_i, S_i^n)$) and has no impact on promotion productivity Y_i^*, then the reference qualification is the promotion productivity and resume absent this line of experience, $Y_i^0 = (Y_i^*, S_i^* \setminus A^n(G_i, S_i^n))$. Section Section 3.3 further discusses the choice of subsystem.

Fix a subsystem $\mathcal{N}^0$ and its corresponding reference qualification. Worker i with group identity G_i enters the subsystem with qualification level Y_i^0. He is evaluated at each node in the subsystem, which (potentially randomly) impacts his productivity and signal at subse-

[8] In a slight abuse of notation, when we write statements across all $n \in \mathcal{N}$, the implied variables for the case of $n = n^*$ are superscripted by n^*, e.g., A^{n^*} corresponds to A^*, etc.

quent nodes. These interactions determine how productivity Y_i^* and signal S_i^*—and hence evaluation A_i^*—at the focal node vary by group. Our definition of total discrimination is with respect to the action distribution at the focal node generated by this process. Formally, let $\sigma^* : \mathcal{Y}^0 \times \{w, b\} \to \Delta(\mathcal{S}^*)$ denote the mapping from qualification and group to signal distribution at node n^*, and let $\alpha^* : \mathcal{Y}^0 \times \{b, w\} \to \Delta(\mathcal{A}^*)$ denote the analogous mapping for the action distribution. The action distribution $\alpha^*(y^0, g)$ for qualification level y^0 and group g can be expressed in terms of the corresponding signal distribution $\sigma^*(y^0, g)$ and the group-g action rule $A^*(g, s)$:

$$\alpha^*(a; y^0, g) = \sigma^*(\{s : A^*(g, s) = a\}; y^0, g). \tag{1}$$

Let $\mu^* : \mathcal{Y}^0 \times \{b, w\} \to \mathbb{R}$ denote the corresponding average action, where $\mu^*(y^0, g)$ is the average action for a group g worker with qualification level y^0. We define total discrimination as the difference between the average actions of group w and b workers with the same qualification level.

Definition 2 (Total Discrimination)**.** *Total discrimination at node n^* for workers with qualification level $y^0 \in \mathcal{Y}^0$ is equal to $\Delta^T(y^0) \equiv \mu^*(y^0, w) - \mu^*(y^0, b)$. Total discrimination arises if $\Delta^T(y^0) \neq 0$ for some $y^0 \in \mathcal{Y}^0$.*

Total discrimination conditions on qualification to capture disparities that arise within the chosen subsystem. In Section 3.3 we discuss how, through the choice of subsystem, our definition of total discrimination nests common notions of discrimination in the literature within a unified framework.

In addition to accounting for how direct discrimination at other nodes contribute to disparities at the focal node, total discrimination includes disparities that arise from direct discrimination at the focal node (since μ^* depends on the action rule, as can be seen in (1) above). *Systemic discrimination* isolates the component of total discrimination that stems from direct discrimination at other nodes by shutting down direct discrimination at the focal node. It does so by considering a counterfactual action distribution where both groups face the same action rule at the focal node. Suppose group g instead faced the group $g' \neq g$ action rule $A^*(g', s)$. The counterfactual action distribution $\tilde{\alpha}^* : \mathcal{Y}^0 \times \{b, w\} \to \Delta(\mathcal{A}^*)$ maps each qualification level y^0 and group g to the action distribution under $A^*(g', s)$ and the group-g signal distribution:

$$\tilde{\alpha}^*(a; y^0, g) \equiv \sigma^*(\{s : A^*(g', s) = a\}; y^0, g). \tag{2}$$

Let $\tilde{\mu}^*(y^0, g)$ denote the expectation of $\tilde{\alpha}^*(y^0, g)$—in other words, the average action at n^* for a group g worker with qualification y^0 if they were to face the action rule for group g' (e.g., group b workers' average action if they faced the group w action rule). Comparing the average counterfactual action of group b workers to the actual average action of group w workers with the same qualification, or vice versa, isolates how decisions at other nodes contribute to disparities at the focal node—that is, systemic discrimination:

Definition 3 (Systemic Discrimination)**.** *Systemic discrimination at node n^* for workers with qualification level y^0 is equal to $\Delta_1^S(y^0) \equiv \mu^*(y^0, w) - \tilde{\mu}^*(y^0, b)$ or $\Delta_2^S(y^0) \equiv \tilde{\mu}^*(y^0, w) - \mu^*(y^0, b)$. Systemic discrimination arises at n^* if $\Delta_1^S(y^0) \neq 0$ or $\Delta_2^S(y^0) \neq 0$ for some $y^0 \in \mathcal{Y}^0$.*

Systemic discrimination also conditions on qualification to capture disparities stemming from systemic forces within the subsystem. See Section 3.4 for a discussion of two key sources of systemic discrimination.

It is straightforward to map this framework to the example in Section 2 studying discrimination in cancer screening recommendations A_i. Economist 1 selects only the doctor's decision node (the focal node), setting the physician's signal as the qualification, $Y_i^0 = S_i$; this includes the nurse practitioner's action. Economist 2 selects the nurse practitioner's and doctor's decision nodes, setting the nurse practitioner's signal as the qualification. Economist 3 selects all nodes that impact information about the patient's cancer status, setting this cancer status as the qualification, $Y_i^0 = Y_i^*$.

To illustrate the definitions of discrimination, consider a Black and white patient who arrive at the nurse practitioner decision node with the same signal (the qualification of Economist 2). Suppose nurse practitioners assign each patient a risk score of low, medium or high (the doctor's signal). Their bias leads the Black patient to receive systematically lower risk scores than the white patient. Suppose doctors underscreen Black patients—specifically, they screen white patients with a medium or high risk score and Black patients with a high risk score. Direct discrimination captures disparities stemming from the doctors' different screening thresholds: white patients with a medium risk score are screened while white patients are not. Systemic discrimination captures disparities stemming from the nurse practitioners' bias in assigning a risk score. Fixing the screening threshold as that used for white patients (i.e., a medium risk score), Black patients are screened at a lower rate, since they are less likely to receive a risk score of medium or high than similar white patients. Total discrimination aggregates both of these components: it compares the screening rate for a white patients who receive risk scores of medium or high to the screening rate of similar Black patients who receive risk scores of high.

This example shows how systemic discrimination arises from the interaction of group-based differences in the signal distribution with the dependence of the action rule on the signal. If instead doctors screened both groups following a medium or high risk score and the risk score distributions differed by race in the probability of low versus medium risk scores but not in the total probability of medium and high risk scores, then Black and white patients are screened at the same rate and there is no systemic discrimination. On the other hand, if the total probability of medium and high risk scores does differ by race, then this difference *does* lead to systemic discrimination.

3.3 Choosing a Subsystem

The study of systemic discrimination requires a researcher to choose what systemic forces to account for via the chosen subsystem. The interpretation of systemic and total discrimination is inherently tied to this choice since only disparities that emerge within the subsystem are included in these measures of discrimination. While prior work often makes this choice implicitly, more explicit discussion is critical for interpreting results and forming an appropriate policy response.

We first map the choice of subsystem $\mathcal{N}^0$ and corresponding qualification Y_i^0 to different common notions of discrimination in the literature. At one extreme, including only the focal node in the subsystem ($\mathcal{N}^0 = \{n^*\}$) corresponds to setting the qualification as the signal at the focal node, $Y_i^0 = S_i^*$. This case does not account for any systemic forces, and total discrimination coincides with direct. This is the choice of Economist 1 in Section 2, and is the implicit choice in most economic analyses of direct discrimination. At the other extreme, including all nodes in the subsystem ($\mathcal{N}^0 = \mathcal{N}$) and setting qualification to a constant, $Y_i^0 = 0$, corresponds to accounting for all systemic forces and assuming there are no differences prior to entry in the system. In this case, total discrimination is the unconditional disparity between groups and systemic discrimination accounts for the impact of all other decisions on productivity and the signal at the focal node. By selecting a set of nodes in between these two extremes, the researcher can isolate different systemic forces in the economy.[9]

Including all nodes that impact the signal S_i^* at the focal node holding fixed productivity at the focal node—which corresponds to setting qualification $Y_i^0 = Y_i^*$—isolates the impact of information on disparities for workers with the same productivity. Total discrimination is then treatment differences for workers with the same productivity. This is the choice of Economist 3, who sets Y_i^0 to the underlying risk of cancer. Notably, this choice aligns total discrimination with the legal notion of *disparate impact*. For example, Arnold et al. (2022), consider a measure of disparate impact in the pretrial setting where qualification is pretrial misconduct potential Y_i^*. This choice also aligns total discrimination with some measures of algorithmic unfairness, where the action is a prediction of some unobserved state Y_i^* (e.g., Berk, Heidari, Jabbari, Kearns, and Roth 2018).

One could also isolate the impact of a single node or subset of nodes. For instance, if S_i^*

[9] See Rose (2022) for a related discussion in the case of direct discrimination. He argues that measuring discrimination—in his case, taste-based or statistical—inherently requires taking a stance on what factors are decision-relevant for the evaluator, and what measures can be classified as discrimination.

is generated by the signal S_i^n and action A_i^n at a prior decision n while Y_i^* is independent of A_i^n, then setting the qualification to be the signal at n and productivity at the focal node, $Y_i^0 = (S_i^n, Y_i^*)$, captures the impact of direct discrimination at node n on treatment at focal node n^* via its impact on signaling for a given level of productivity. For example, suppose a club membership acts solely as a signaling device and has no impact on productivity. By including this node, total discrimination accounts for the signaling impact of differential access to club membership.[10] This is the choice of Economist 2 in Section 2, who sets Y_i^0 to be the information available to the nurse practitioner. This accounts for the impact of direct discrimination in the nurse practitioner's decision on the doctor's information, and hence, decision.

Similarly, one could include a node that impacts productivity at the focal node. For instance, if Y_i^* is generated by the productivity Y_i^n and action A_i^n at a prior decision n, and conditional on Y_i^* both groups generate the same distribution of signal S_i^*, then setting the qualification to be productivity at this prior node, $Y_i^0 = Y_i^n$, captures the impact of direct discrimination at n on treatment at n^* via the induced disparities in Y_i^*. For example, direct discrimination in prior employment may generate disparities in current productivity. Including this prior employment node accounts for the systemic impact of this prior discrimination.

In addition to including nodes where decisions are made prior to the focal node, a researcher may also include nodes at which decisions are made contemporaneously with or subsequent to the focal node. Here, the anticipation of direct discrimination at these nodes may influence the signal and productivity at the focal node. For example, anticipating a discriminatory jury may impact a defense attorney's decision to accept a plea deal. Here, the plea deal decision is the focal node n^*, the jury decision is the other node n, and the anticipated action $A^n(G_i, S_i^n)$ is part of the signal S_i^* at the focal node, and hence, enters $A^*(G_i, S_i^*)$.

Thus, through the choice of subsystem, Definitions 1 to 3 provide a unified framework for studying different notions of discrimination in the literature and, as in the motivating example, can help interpret and integrate seemingly disparate findings. Considering different subsystems also allows for the study of policy interventions at different decision nodes.

An open debate is whether group-based differences in preferences that generate differences in productivity or the signal should be coded as discrimination. The subsystem can be chosen to include or exclude such preference differences through the inclusion or exclusion of nodes where these preferences influence actions. For example, suppose racial or gender socialization affects the worker's decisions in a way that affects her work history or ability to

[10] This is the reasoning behind legal cases made against group-based exclusivity in country clubs, which offer members a host of pecuniary and non-pecuniary benefits such as access to networks (Jolly-Ryan 1998).

signal her productivity (e.g., choosing a job with a flexible schedule, refraining from asking for a raise). Including nodes where the worker makes these choices accounts for this as systemic discrimination, while excluding such nodes does not (as in, e.g., Cook, Diamond, Hall, List, and Oyer 2021).

3.4 Discussion

Relation to Systemic Discrimination in Other Literatures. Our definition of systemic discrimination aligns broadly with how systemic and structural discrimination are discussed in the sociology literature: as a form of inequality operating indirectly through characteristics beyond group identity and stemming from discrimination in other parts of the system. Pincus (1996) defines structural discrimination as referring to "the policies of dominant race/ethnic/gender institutions and the behavior of individuals who implement these policies and control these institutions, which are race/ethnic/gender neutral in intent but which have a differential and/or harmful effect on minority race/ethnic/gender groups" (see also Hill (1988)).[11] Correspondingly, in our definition, systemic discrimination can generate total discrimination even when there is no direct discrimination at the focal node—i.e., the action rule $A^*(g, s)$ is group-neutral—because this group-neutral action rule fails to account for discrimination at other nodes or even intentionally builds in discrimination indirectly by using other information to proxy group.[12] Powell (2007) defines systemic discrimination as a "product of reciprocal and mutual interactions within and between institutions," both "within and across domains."[13] Similarly, our definition of systemic discrimination captures disparities that arise from the interaction between discriminatory decisions across time and contemporaneously across different domains. In our setting, systemic discrimination can emerge when past discriminatory decisions impact present decisions—so-called "past-in-present" discrimination (Feagin and Feagin 1978), as illustrated in Sections 5 and 6.[14]

[11] For example, the historical practice of "redlining" in mortgage markets prioritized borrowers from majority-white neighborhoods over equally-creditworthy borrowers from majority-Black neighborhoods. Such neighborhood-based prioritization generated substantial race-based lending disparities despite the policy being prima facie race-neutral.

[12] Our definition also aligns broadly with some notions of institutional discrimination (Small and Pager 2020), though other forms, e.g., when direct discrimination is codified into policy such as the case of Jim Crow laws, is a separate phenomenon.

[13] He terms discrimination arising from the interactions of systems as "structural" and discrimination stemming from interactions in a system as "systemic." We do not formalize this distinction here, but it follows naturally from our framework.

[14] Past-in-present discrimination can also emerge when a system or institution is first "designed" by a group in power, which leads to the development of evaluation criteria that are optimized around the characteristics of this group. For example, De Plevitz (2007) discusses the impact of the "Eurocentric model of teaching" on schooling outcomes of Aboriginal children in Australia, noting that by not accounting for the family structure and cultural obligations of the Aboriginal community the educational system creates systemic barriers for the minority population. Another example is the practice of excluding women or minority groups from medical trials, which leads to a less informative signal of the efficacy of new treatments for these groups (Bierer, Meloney, Ahmed, and White 2022). In our framework, this corresponds to viewing the signal distribution

It can also can emerge across domains when discriminatory practices in one market impact productivity or signaling in another—so-called “side-effect” discrimination (Feagin and Feagin 1978), as illustrated in ??. We further review connections to the sociology literature on systemic discrimination, as well as notions of discrimination in law, economics, and computer science, in Appendix A.

Sources of Systemic Discrimination To delineate two sources of systemic discrimination, we split group-based differences in the signal distribution $\sigma^g(s; y^0)$ into two components: an *informational* channel stemming from group differences in the signal distribution $\sigma^g(s|y^*; y^0)$ for workers with the same productivity and qualification level, and a *technological* channel stemming from group differences in the productivity distribution $\phi^g(y^*; y^0)$ for workers with the same qualification level.[15] We discuss each in turn.

Informational systemic discrimination emerges from group-based differences in how signals are generated among workers who are equally productive at the task at hand. One salient form of informational systemic discrimination is *signal inflation*, in which for a given level of productivity, the signal is on average higher for one group than the other, and higher signal realizations lead to more favorable actions. For example, Black defendants with the same potential for pretrial misconduct (Y_i^*) as white defendants are less likely to have a clear criminal record (S_i), which decreases their probability of being released on bail (Pager, Bonikowski, and Western 2009; Agan and Starr 2017). Such signal inflation can arise from direct discrimination in other interactions with law enforcement.[16] Since signal inflation is a statistical bias in the productivity signal, it can be offset by an action rule that corrects for it, i.e., via “reverse” direct discrimination. For example, if direct discrimination in policing leads to criminal record disparities and the bail judge is aware of this past discrimination, then she can account for it in her interpretation of criminal records.

Another salient form is *differential screening*, where the manager has a more precise signal for one group than the other (Cornell and Welch 1996).[17] For example, a test is trained to screen men and generates less reliable information about the productivity of

as a choice variable for the dominant group, similar to the discussion in Pincus (1996).

[15] Note that the full signal distribution can be constructed from these two components, $\sigma^g(s; y^0) = \int_{\mathcal{Y}^*} \sigma^g(s|y^*; y^0)\phi^g(y^*; y^0)dy^*$.

[16] For example, Pierson, Simoiu, Overgoor, Corbett-Davies, Jenson, Shoemaker, Ramachandran, Barghouty, Phillips, Shroff et al. (2020) show that Black individuals are more likely to be stopped by police and charged with a crime. Other examples of signal inflation include recruitment practices that prioritize workers with certain social connections, where one group is more connected than equally qualified members of the other, and wage setting based on salary history, where one group has higher past salaries than equally productive members of the other group. For example, Agan, Cowgill, and Gee (2021) study how salary disclosure impacts wage offers. They find little evidence for direct discrimination conditional on a given salary disclosure, but sizeable treatment disparities stemming from the lower disclosed salaries for women.

[17] In Cornell and Welch (1996), minority job applicants receive fewer draws of a binary signal than majority applicants. This is a form of systemic discrimination, as if minority and majority applicants received the same number of signal draws, they would be evaluated equally (i.e., there is no direct discrimination).

women (Mocanu 2022).[18] Another example is borrowers with the same ability to repay (Y_i^*) have differentially informative credit histories (S_i) due to discrimination in past borrowing opportunities, which provide opportunities to signal creditworthiness (Bartik and Nelson 2016). Unlike signal inflation, differential screening cannot be offset by the manager's action rule; to eliminate it, the manager needs to collect more precise information for one group (or ignore the more precise information for the other group). Note that differential screening will also lead to direct discrimination when the signal precision impacts the action rule, as shown in (Aigner and Cain 1977). Appendix B illustrates how differential screening can lead to both direct and systemic discrimination.

Technological systemic discrimination emerges from group-based differences in productivity at the focal node for workers who enter the subsystem with the same qualification. It can arise when Y_i^* is systematically higher for members of one group, holding fixed Y_i^0. For example, white workers might have more access to training and skill development than Black workers due to discrimination in education and the labor market, and hence, have more opportunities to build human capital.[19] Alternatively, Black workers may respond to anticipated future direct discrimination by investing less in human capital (Coate and Loury 1993).[20] Technological systemic discrimination also includes the type of "task-based" discrimination studied in Hurst et al. (2021), where workers have no initial group-based differences but racial barriers to specialization generate group-based differences in which tasks a worker chooses to specialize in (Y_i^*).

The chosen subsystem impacts the potential sources of systemic discrimination. At the one extreme, when all nodes are included and the qualification is set to a constant, all differences in the signal and productivity distributions contribute to systemic discrimination. At the other extreme, when only nodes that impact signaling are included and qualification is set to productivity, all systemic discrimination is informational. In between these two extremes, informational and technological channels can both contribute to systemic discrimination.

Relation to Statistical (Direct) Discrimination Statistical (direct) discrimination also stems from group-based differences in the signal and productivity distributions, but it conceptually differs from systemic discrimination. Such statistical discrimination arises from the impact of the signal and productivity distributions on the action rule; in contrast, systemic

[18] Specifically, they show that subjective tests designed to screen men led to disparate outcomes for women; amending or replacing the tests with more objective evaluations mitigated disparities. Pinkston (2003) also finds evidence that employers receive less-accurate initial signals from women than from men and De Plevitz (2007) document disparities arising from using height-to-weight ratios calibrated with Anglo-Celtic data in job screening.

[19] Gallen and Wasserman (2021) provide evidence for this channel via gender differences in career advice, where women are more likely to receive advice about work/life balance than men. This can deter investment in human capital and the pursuit of careers in competitive fields.

[20] Here, workers have no group-based differences in initial productivity (Y_i^0). They make a costly decision to invest in human capital that increases productivity at the focal node (Y_i^*).

discrimination arises from the impact of these distributions on the action *distribution* for a given qualification level. When $Y_i^0 \neq S_i$, differences in the signal distribution can lead to both systemic discrimination and statistical direct discrimination. Similarly, when $Y_i^0 \notin \{Y_i^*, S_i\}$, differences in the productivity distribution can lead to systemic discrimination and statistical direct discrimination. In both cases, focusing only on direct discrimination would miss a key aspect of how group differences in the signal and productivity distributions contribute to action disparities.

Finally, we note that unlike with statistical discrimination (e.g. Bordalo, Coffman, Gennaioli, and Shleifer 2019; Bohren et al. 2022) there is no scope for "inaccurate" systemic discrimination: only the true productivity and signal distributions contribute to systemic discrimination. However, inaccurate beliefs about the signal and productivity distributions can lead to inaccurate perceptions about the extent of systemic discrimination. This can affect action choices, and hence, total discrimination. It can also impact the choice of signaling technology. For example, a mortgage assessor may incorrectly believe that a particular credit score provides an identical signal of creditworthiness across groups, and therefore continue using it without adjusting for group differences.

3.5 Decomposing Total Discrimination

We next connect these three definitions by showing how total discrimination can be decomposed into direct and systemic components. From Definition 2, $\Delta^T(y^0)$ measures average total discrimination at qualification level $y^0 \in \mathcal{Y}^0$ and $\Delta_1^S(y^0)$ or $\Delta_2^S(y^0)$ measures average systemic discrimination at y^0. A measure of direct discrimination at signal realization $s \in \mathcal{S}$ is given by the difference between the selected actions for group w and group b, $\tau(s) \equiv A(w,s) - A(b,s)$. *Average direct discrimination* at qualification level $y^0 \in \mathcal{Y}^0$ is the expected direct discrimination with respect to the signal distribution for group g at qualification y^0:

$$\overline{\tau}(g, y^0) \equiv E[\tau(S_i) \mid G_i = g, Y_i^0 = y^0], \tag{3}$$

for $g \in \{w, b\}$. While each of these measures are for a particular qualification level, it is also possible to construct an overall measure by averaging across qualification levels.[21]

Our decomposition expresses total discrimination at qualification level y^0 as the sum of two terms: average direct discrimination with respect to the signal distribution for group w workers with qualification level y^0 and systemic discrimination at qualification level y^0 when

[21] The interpretation of this overall measure depends on the chosen qualification distribution: averaging across the population qualification distribution yields a measure of average discrimination across both groups, while averaging across the qualification distribution for group g yields a measure of average discrimination for a group g worker.

the manager uses the action rule for group b,

$$\underbrace{\mathit{\Delta}^T(y^0)}_{\text{Total discrimination}} = \underbrace{\overline{\tau}(w,y^0)}_{\text{Avg. direct discrimination}} + \underbrace{\mathit{\Delta}_2^S(y^0)}_{\text{Systemic discrimination}} . \tag{4}$$

This is in the spirit of Kitagawa (1955); Oaxaca (1973); Blinder (1973), who relate unconditional disparities to a component explained by observable worker characteristics (e.g., education or labor market experience) and a residual "unexplained" disparity. These classic decompositions can be viewed as a strategy for measuring direct discrimination, which attempts to hold fixed all relevant non-group characteristics. Equation (4), in contrast, leads to strategies (developed below) for measuring systemic discrimination as the residual of total discrimination after accounting for direct discrimination.

As with the classic Kitagawa-Oaxaca-Blinder approach, there are multiple ways to decompose total discrimination into direct and systemic components, and the "order" of the decomposition may matter empirically. In particular, we can also express total discrimination as the sum of average direct discrimination with respect to the signal distribution for workers from group b and systemic discrimination when the firm uses the action rule for group w, all at qualification level y^0:

$$\mathit{\Delta}^T(y^0) = \overline{\tau}(b,y^0) + \mathit{\Delta}_1^S(y^0). \tag{5}$$

Finally, averaging (4) and (5) yields a third decomposition:

$$\mathit{\Delta}^T(y^0) = \overline{\tau}(y^0) + \overline{\mathit{\Delta}}^S(y^0), \tag{6}$$

where $\overline{\mathit{\Delta}}^S(y^0) \equiv \frac{1}{2}(\mathit{\Delta}_1^S(y^0)+\mathit{\Delta}_2^S(y^0))$ averages the systemic discrimination terms and, slightly abusing notation, $\overline{\tau}(y^0) \equiv \frac{1}{2}(\overline{\tau}(w,y^0)+\overline{\tau}(b,y^0))$ averages the direct discrimination terms.

Each of these three decompositions yield a measure of systemic discrimination as the difference between total discrimination and the average direct discrimination component. The challenge of identifying systemic discrimination thus reduces to the challenge of measuring average direct and total discrimination. We next discuss different identification strategies.

4 Identification

We next show how the above framework can be brought to data with the iterated audit (IA) approach. To formalize this approach simply, consider a two-node system with the reference point chosen as the initial-node signal, $Y_i^0 = S_i^1$, which we assume is observed by the econometrician. We discuss at the end of this section how the approach extends to multi-node systems and other choices of qualification.

4.1 Treatment and Interaction Components

Measuring systemic discrimination generally involves the measurement of two conceptually distinct components: *(i)* a treatment component capturing direct discrimination at focal and non-focal nodes and (*ii*) an interaction component capturing how the actions at non-focal nodes impact the action at the focal node via focal-node signals. Direct discrimination, and hence the treatment component, can be identified by conventional experimental designs such as a standard audit or correspondence study. The key challenge to measuring systemic discrimination is therefore identification of the interaction component.

To formalize the identification challenge, recall that systemic discrimination is measured by $\Delta_1^S(y_0) = \mu^*(y^0, w) - \tilde{\mu}^*(y^0, b)$ or $\Delta_2^S(y_0) = \tilde{\mu}^*(y^0, w) - \mu^*(y^0, b)$. In our two-node system with $Y_i^0 = S_i^1$, the key objects in these measures can be written:

$$\begin{aligned} \mu^*(y^0, g) &= E[A^*(g, S_i^*) \mid Y_i^0 = y^0, G_i = g] \\ &= A^*(g, S^*(A^1(g, y^0), y^0)) \end{aligned} \tag{7}$$

$$\begin{aligned} \tilde{\mu}^*(y^0, g) &= E[A^*(g', S_i^*) \mid Y_i^0 = y^0 G_i = g] \\ &= A^*(g', S^*(A^1(g, y^0), y^0)), \end{aligned} \tag{8}$$

recalling that $A^*(g, s)$ and $A^1(g, s)$ are the action rules at the focal and initial node, respectively. Here we also define $S^*(a, y^0)$ as the focal-node signal that is realized given an initial-node action of a among those with $Y_i^0 = y^0$, such that $S_i^* = S^*(A_i^1, S_i^1)$.

Clearly, knowing how group membership directly affects actions at the focal and initial nodes (formally, how $A^*(g, s)$ and $A^1(g, s)$ depend on g) is not enough to identify equations (7) and (8). It is also necessary to know how initial-node actions indirectly affect focal-node actions through focal-node signals (formally, how $A^*(g, S^*(a, y^0), y^0)$ depends on a). The former is what we term knowledge of the treatment component; the latter is what we term knowledge of the interaction component.

To make Equations (7) and (8) concrete, consider a two-node labor market setting in which workers first apply for an internship (the initial node) and then apply for an entry-level position (the focal node). Equations (7) and (8) show that systemic discrimination in entry-level hiring actions arises from two distinct components: a treatment component capturing direct discrimination in internship and entry-level hiring and an interaction component capturing how internship experience impacts entry-level hiring. To estimate the former treatment component, a researcher could run a conventional audit study where workers of different races or genders apply to internships with identical resumes. But this audit would not identify the interaction component and hence not identify systemic discrimination.

We develop two alternative iterated audit approaches to estimating the interaction component: a constructive approach which separately estimates the impact of each possible action at other nodes on the focal node action, and an experimental approach which directly measures the interaction component by simulating impacts of different action distributions at other nodes. Note that while treatment-component knowledge is enough to identify direct discrimination at both the initial and focal nodes, knowledge of both components is also generally needed to identify total discrimination (i.e. $\mathit{\Delta}^T(y^0) = \mu^*(y^0, g) - \mu^*(y^0, g'))$ as well as the discrimination decompositions (4), (5) and (6).

4.2 The Constructive IA Approach

The first IA approach separately estimates the impact of each possible action at non-focal nodes on focal-node signals, as well as the effect of each possible focal-node signal on focal-node actions. In our two-node example, this would mean learning how $S^*(a, y^0)$ depends on a as well as how $A^*(g, s)$ depends on s. With these estimates, a researcher can *construct* an estimate of the interaction component: e.g. how $A^*(g, S^*(a, y^0), y^0)$ depends on a.

To make this constructive IA approach concrete, consider again the entry-level job hiring example. A researcher could first learn how past internship experience translates to the sets of signals a hiring manager sees when evaluating candidates for the entry-level job. For example, she could see that candidates with internship experience list it on their resumes which hiring managers evaluate. The researcher could then learn how hiring manager actions depend on these signals; e.g. she could randomize different internship experiences to resumes and measure hiring decisions. Combining these steps, the researcher can construct a measure of how internship experience impacts entry-level hiring. Combining this with conventional audit information on direct discrimination in both internship and entry-level hiring, she could then measure systemic (and total) discrimination in entry-level hiring.

This example also illustrates the importance of measuring the interaction component; i.e., why conventional audit study estimates of the treatment component are not enough to identify systemic or total discrimination in entry-level hiring. Even if direct discrimination in both internship and entry-level hiring is minimal, whether or not an entry-level applicant has previous internship experience could be critical for their success in hiring. In this case, minimal direct discrimination in internship hiring could lead to large systemic and total discrimination in entry-level hiring through the interaction component. More generally, nonlinearities in how actions at non-focal nodes affect focal-node actions can mean that conventional audit study analyses of direct discrimination at different nodes fail to reveal the total extent of discrimination through interactions across nodes. Conversely, decision-makers might act to undo perceived direct discrimination in the signals they observe, such that discrimination at non-focal nodes does not translate to total discrimination at the focal node. Entry-level hiring managers, for example, could put less weight on internship experience for groups that have historically faced discrimination in such opportunities. In this case, even if a conventional audit reveals sizable direct discrimination in internship hiring, total discrimination in entry-level hiring may be minimal.

While conceptually straightforward, the constructive IA approach will likely prove challenging to implement when action or signal spaces are high-dimensional or otherwise complex. Suppose, for example, that workers can be hired for a wide range of internship positions with different responsibilities and tasks, and that information about these internship experiences are signalled to entry-level hiring managers by a free response text box. Here the set of possible initial-node actions is large, making it challenging to estimate how $S^*(a, y^0)$ depends on a. The unstructured text data making up entry-level hiring manager signals further makes it difficult to estimate how $A^*(g, s)$ depends on s. To address these practical issues, we turn to the experimental IA approach.

4.3 The Experimental IA Approach

The second IA approach directly measures systemic discrimination by simulating the distribution of focal-node signals, as impacted by group membership through the actions at other nodes in the subsystem, and measuring focal-node actions given the simulated signals. In our two-node example, this would mean generating draws of $\tilde{S}_i^* \mid G_i = g, Y_i^0 = y^0$ and measuring $A^*(g, \tilde{S}_i^*)$ and $A^*(g', \tilde{S}_i^*)$ to *experimentally* measure $\mu^*(y^0, g)$ and $\tilde{\mu}^*(y^0, g)$.

To make this experimental IA approach concrete, consider the previous example with entry-level hiring managers observing unstructured text data about an applicant's possible internship experience. A researcher could obtain a set of internship experience descriptions (including no description, for those without internships) from individuals of different groups g. She could then generate entry-level job applications by drawing descriptions from this distribution and attaching a salient signal of group membership—both the same g or a different group g'. The average entry-level hiring manager action given these simulated applications identifies either $\mu^*(y^0, g)$ or $\tilde{\mu}^*(y^0, g)$, and hence systemic (and total) discrimination.

While being more practical in settings with high-dimensional or otherwise complex signals, this approach has the drawback of not separately measuring the impact of initial-node actions on focal-node signals as in the constructive IA approach. Consequently, the precise mechanisms for systemic discrimination may be harder to infer with the experimental IA approach. Still, by capturing the interaction component, the experimental IA approach will again capture any nonlinearities or offsetting behavior that might obscure systemic discrimination from otherwise accumulated estimates of direct discrimination across nodes.

5 Constructive Iterated Audit

Our first study demonstrates the constructive IA approach.[22] We use a field experiment to study a system where applicants apply for a job at two nodes. At the first node, applicants apply with no prior work experience and then get hired or not. They apply at the second

[22] See https://aspredicted.org/5XT_YPY for pre-registration.

(focal) node either having obtained prior experience or not. This setting is similar in spirit to the example outlined in Section 4.2. Direct discrimination in hiring at the first node can thus generate systemic discrimination at the second node due to race-based differences in experience. Our experiment allows us to separately measure the impact of each possible action at the first node on treatment at the focal node. We then use these measures to estimate the treatment and interactive components and show how these can be used to capture total, direct, and systemic discrimination in our setting.[23]

5.1 Method

The experiment builds on the correspondence study design (Bertrand and Mullainathan 2004b; Kline et al. 2021) by generating sets of fictitious resumes that varied in their race and non-race attributes. Race was varied through the applicant's name, which was either stereotypically White or Black, and experience was varied by whether the applicant had a line of previous experience in the same sector and geographic (metro) area as the job being applied for. These resumes were submitted to online job vacancies at a pre-determined group of national firms. Our main measure of interest was the call-back rate for each set of resumes. We targeted automotive firms because this sector was shown to have the largest racial gap in call-back rates in prior work (Kline et al. 2021).[24] All experimental details can be found in Appendix B.

We study a subsystem with two nodes, $\mathcal{N}^0 = \{1, n^*\}$: applying for a job with no prior experience (Node 1) and applying for a job after potentially gaining previous experience (Node n^*, the focal node). Here, $A^1(G_i, S_i^1)$ at Node 1 corresponds to the decision to hire or not hire worker i from group G_i with no previous work experience. This action generates a signal $S_i^{n^*}$ at the focal node n^* which corresponds to whether the worker obtained previous experience or not. $A^{n^*}(G_i, S_i^{n^*})$ at Node n^* corresponds to the hiring decision at the focal node, which is a function of the action rule at Node 1. This setting is ideal for the constructive IA approach because it is straightforward to estimate the impact of each possible action at the non-focal on focal node signals.

The key aspect of the IA method is the capacity to measure total and systemic discrimination through the estimation of the treatment and interaction components by keeping the reference qualification Y_i^0 fixed and constant across workers. In the Constructive IA design, estimating the interaction component requires (i) assessing how direct discrimination at Node 1 translates to race-based differences in in the signal (experience) at the focal node n^* and then (ii) estimating the return to experience at n^*. To estimate how direct discrimination at Node 1 leads to disparities in experience (i), we submitted otherwise identical

[23] See Appendix B.2 for detailed instructions.

[24] We used the same set of automotive firms as those targeted in (Kline et al. 2021). Notably, our experiment was completed prior to the public release of those firms in Kline, Rose, and Walters (2024).

resumes with no previous experience which differed only in the race of the applicant. To estimate returns to experience (ii), we submitted otherwise identical resumes that differed in the race of the applicant and included a line of previous entry-level experience at a related job.

5.1.1 Local Market Thickness and Call-Back Conversion

A difference in direct discrimination across nodes, combined to returns from experience, can potentially generate significant systemic discrimination in the system. However, the extent of systemic discrimination depends critically on two additional factors: *local market thickness*—i.e., the number of local available jobs—and the prospective call-back conversion rate—i.e., the proportion of call-backs that result in a job offer. Intuitively, when the number of available jobs is very low, direct discrimination as measured by the call back rate is the primary statistic relevant for disparities; workers with or without previous experience have few jobs to apply for after getting rejected. On the other hand, when the number of jobs is very high, the call back rate will not be very relevant since applicants will end up landing a job with enough applications. Systemic discrimination plays a larger role at intermediate levels of market thickness since the discriminated group is less likely to obtain credentials (experience) associated with higher propensities of landing a job. The call-back conversion rate is important because total discrimination is a function of disparities in obtaining previous experience rather than call-backs per se.

To measure each type of discrimination, we collected two additional pieces of data. First, we scraped local job openings in automotive firms over time across all of the municipalities targeted in our experiment; we use this measure as a proxy for local market thickness in the target industry.[25] The 25^{th}, 50^{th}, and 75^{th} percentiles of job openings within the multiplicities were 5, 8, and 18, respectively. Second, We used a hiring and recruitment agency to recruit hiring managers ($N = 107$) with experience in evaluating applicants to entry-level jobs in the automotive industry to obtain a measure of the conversion rate for call-backs at each node.[26] Hiring managers were surveyed on the proportion of call-backs that would be converted to a job offer as a function of previous experience and race. Conditional on a call-back, the race-based difference in job offers was small and insignificant. However, there was a significant gap in the job offer conversion rate as a function of experience: 55% versus 71% of call-backs to those without versus those with experience, respectively ($p < .01$).

[25] See B.1.4 for details about the procedure.

[26] See B.1.5 for details about the procedures.

5.1.2 Call Analysis

We considered first-time phone calls to each applicant from each company. We counted the phone calls for each candidate as follows. First, monitoring from Google Voice, we identified phone numbers that were associated with the companies we applied to by examining their displayed names on Google Voice (e.g., some numbers will show the name of the company) and examining their voicemail, especially the companies they are from. At this step, we have a dictionary of verified phone numbers and their companies. Second, we filtered phone calls from these verified numbers, and mapped phone calls from these numbers to our fictitious applicants based on the recipient (applicant) phone number. In the cases where the same company calls the same applicant multiple times, we only record the first call from this company. Here we have a dataset of applicant-company pairs where there was at least one call. Lastly, we mapped all the applicant-company pairs to the original datasets of applicants, to map the first-time calls to the treatments based on the applicant-company pair. We then summarized across the treatment arms to arrive at the final estimates of first-time calls to each treatment arm.

5.2 Results

We begin by looking at the call-back differences by the race and experience level of the applicants. Looking at those without experience, 25.6% of White resumes received a call back but only 13.2% of Black resumes did so. This 12.4 percentage point racial gap (94% increase) corresponds to significant direct discrimination at Node 1 of our sub-system ($p < .01$).[27] At the same time, we observe substantial returns to prior experience. Pooling across race, resumes with prior experience received 9.5% more call-backs (49% increase) than those without previous experience ($p < .01$). Looking at direct discrimination for those with experience at the focal node $n*$, 31% of White resumes and 27% of Black resumes received a call back. Experience shrunk the racial gap in call-back rates to 4 percentage points, which corresponds to only a 15% difference. This disparity roughly matches the gap in Kline et al. (2021), who also look at experienced applicants. 1 presents these results while controlling for city fixed effects.

The combination of data on direct discrimination at the two nodes, as well as data on call-back conversion rates and local market thickness, can then be used to measure total and systemic discrimination at the focal node n^*. B.1.6 outlines the method of calculating each measure of discrimination as a function of the latter two factors. 1 shows the extent of total, systemic, and direct discrimination as a function of market thickness. From the figure, one can readily see the intuition that when there are very few jobs, direct discrimination constitutes the vast majority of total discrimination. As the number of jobs increases, the

[27] p-values calculated from two-sample t-tests.

Table 1. Call-Back Rates by Racial Group and Experience

Variable	(1)	(2)	(3)
Minority	-0.094***		-0.133***
	(.023)		(0.031)
Experience		0.088***	0.045
		(0.022)	(0.031)
Minority*Experience			0.083*
			(0.044)
City Fixed Effects	Y	Y	Y
Observations	1,001	1,001	1,001
R-squared	0.532	0.530	0.543

Note: Standard errors in parentheses, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

disparity in obtaining prior experience—stemming from direct discrimination at the initial Node 1—starts to play a larger role. Given the high returns to experience, systemic discrimination begins to constitute a higher proportion of total discrimination at the focal node n^*. In fact, looking at the median local market thickness in our experiment, direct discrimination misses nearly 50% of the total discrimination. Namely, the level of total discrimination at the focal node n^* is nearly *double* the level of direct discrimination, with the gap being driven by systemic discrimination.

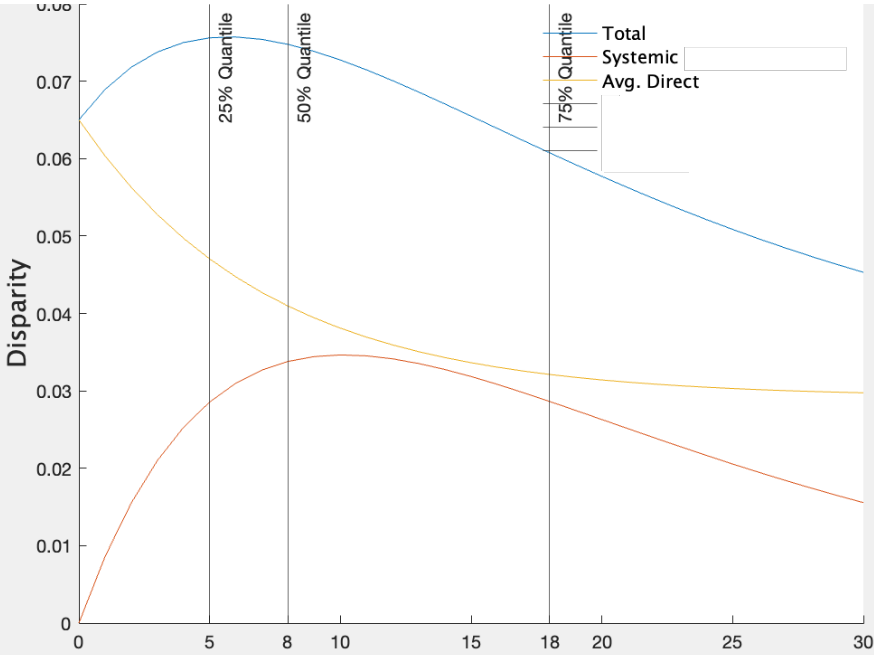
Notably, this analysis also highlights the role of local market thickness in targeting potential policy interventions. Our results suggest that systemic discrimination stemming from disparities in initial, entry-level hiring is particularly significant in markets with intermediate levels of thickness. In those cases, it becomes particularly important to target interventions to earlier parts of the pipeline where workers are just starting out.

6 Experimental Iterated Audit

While the constructive IA approach works for identifying the different components of discrimination when the data is relatively low-dimensional, e.g., binary data such as call-back rates, it is difficult to use in settings with high-dimensional or otherwise complex data. In these types of settings the Experimental IA approach can be used. We now proceed to illustrate the Experimental IA approach in a labor market context where evaluations require interpreting high-dimensional attributes in the form of text data.footnoteSee https://aspredicted.org/w4g6-kzgq.pdf for pre-registration.

6.1 Method

Our second study illustrates the value of the IA design when the non-group signal is high-dimensional or complex. Recent work has argued that direct discrimination is unlikely to



Note: Disparity as a function of local market thickness.

FIGURE 1. Constructive Iterated Audit: Decomposition

explain gender disparities in the labor market. For example, a meta analysis of correspondence studies has found little support for discrimination when the non-group information of male and female candidates is held fixed (Schaerer et al. 2023). At the same time, studies have found significant disparities in how male and female candidates are described to potential employers—the recommendation letters written for similarly qualified male versus female candidates differ significantly in the language used (Schmader et al. 2007; Trix and Psenka 2003). How do such differences in language contribute to disparities? The answer is not obvious. For example, hiring managers may focus on 'hard' information contained in the resume and ignore the differences in language used; alternatively, the differences may lay in language that is not relevant for the evaluation decision.

Given the high-dimensionality of language data, we examine this question through the experimental IA approach which aims to measure the role of disparities in language in generating systemic discrimination in labor market outcomes. We again consider a subsystem with two nodes $\mathcal{N}^0 = \{1, n^*\}$: labor market candidates interact with recommendation letter writers (Node 1) and then present their qualifications, along with recommendation letters, to hiring managers at the focal node (Node n^*, the focal node). In this setting, $A^1(G_i, S_i^1)$ at Node 1 corresponds to the choice of what language to use when writing a letter for worker i from group G_i. The action A^1 generates a high-dimensional signal $S_i^{n^*}$ at the focal node n^*, which corresponds to the reference letter. Finally, as in the first study, $A^{n^*}(G_i, S_i^{n^*})$ at Node n^* corresponds to the hiring decision at the focal node, which is a function of the type of letter generated by the action rule at Node 1. Given the high dimensionality of the signal $S_i^{n^*}$, it is not feasible to measure the impact of each possible action at Node 1 on actions at n^*. The experimental IA is ideal for these types of settings.

Prior work documenting language differences in recommendation letters used existing letters written for different candidates. Since these letters mention different qualifications and experience, it is not possible to use them while ensuring that Y^0 is held fixed across male and female candidates. To get around this issue, we leverage recent work showing that prompting Large Language Models with the same information about candidates—while only varying the gender signaled by the name—reproduces the language differences in recommendation letters found in observational studies (Wan et al. 2023). We replicate this approach by providing an LLM with otherwise identical resumes for male and female candidates to STEM jobs that typically require recommendation letters and prompting them to generate a set of recommendation letters. Note that keeping resumes fixed allows us to test for the impact of disparities in language while keeping qualification fixed.[28] We observe similar significant gender differences in language on dimensions such as individual agency, leadership ability, and communality as in prior work (Wan et al. 2023; Trix and Psenka 2003; Schmader et al. 2007).[29] These sets of recommendation letters constitute the signals seen by evaluators at the focal node $S_i^{n^*}$.

We employ a lab-in-the-field experiment to measure how the gender-based disparities in signals translate to disparities in the action rule at n^*. We used a hiring and recruitment agency to recruit hiring managers ($N = 296$) with experience in recruiting for STEM jobs that typically require a recommendation letter and who were currently looking for employees. Each hiring manager evaluated a random draw of four fictitious candidates on the likelihood that they would recommend the applicant to the next stage of the recruitment process (scale of 1 to 10) and on their expected (hourly) wage should they be hired. Each candidate came with a set of materials composed of a resume and a recommendation letter. Decisions were incentivized using a similar methodology to Kessler et al. (2019): the resumes themselves were fictitious, but the components (e.g., prior work experience) could be matched to resumes of actual potential applicants who had similar attributes and presented to the managers based on their likelihood scores.[30]

The experimental IA design featured three sets of materials, as depicted in Figure 2. For each candidate, a set of materials consisted of a resume and a recommendation letter. Two of the three sets, referred to as Endogenous-m (A) and Exogenous-f (B), were similar to those

[28] See Appendix B.2 for details on how these letters were generated.

[29] For example, "Matthew independently spearheaded projects to create..." versus "Emily reliably developed and maintained software applications..."; "Jacob Meyer is a self-driven, highly capable professional" versus "Mary is a diligent professional with strong communication skills..."

[30] This factorial design is known as an Incentivized Resume Rating paradigm. See Lahey and Oxley (2021) and Kübler, Schmid, and Stüber (2018) for similar uses of factorial designs in studying discrimination.

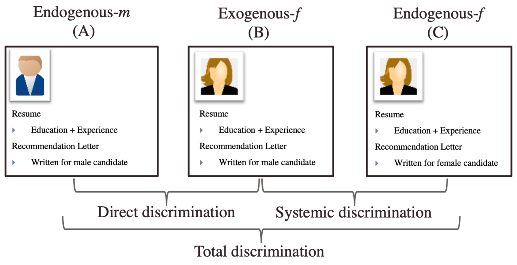


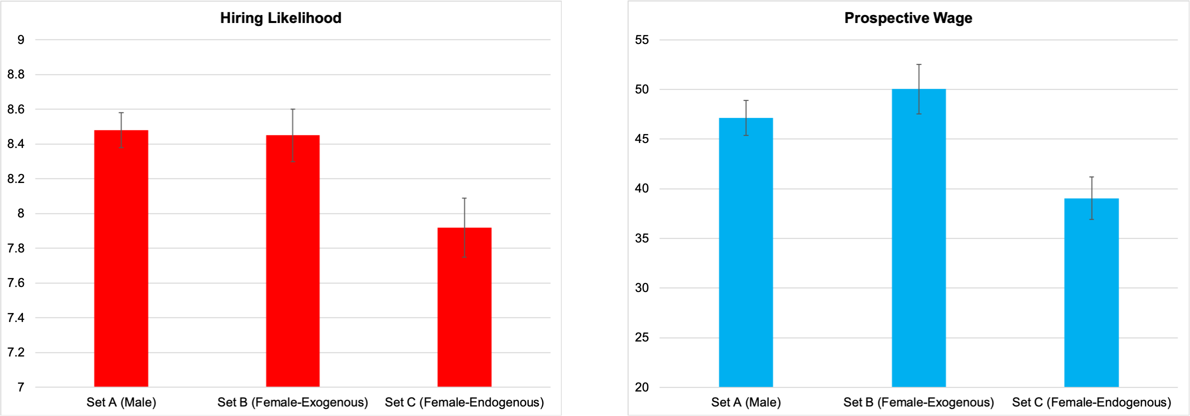
FIGURE 2. Experimental Iterated Audit

in a standard correspondence or audit study: the resumes and recommendation letters were the same aside from perceived group identity (in our case, distinctively male or female name). The resumes differed only in the name of the candidate and the recommendation letters were prompted to be written for a male candidate, with the name and relevant pronouns switched to female in the case of set B. The third set of materials, Endogenous-f (C), had female perceived group identity (distinctively female names) but differed from set (B) in that the recommendation letters were written for a female candidate.

6.2 Results

Our results show a significant impact of language on driving total and systemic discrimination at the focal node. The experimental IA design allows us to identify the measures of total, systemic and direct discrimination by a simple comparison of means. Figure 3 presents this breakdown graphically. Comparing evaluations of materials in sets A and C identifies total discrimination: candidates in set C had a substantially lower hiring likelihood and prospective wages—roughly 28.6% and 31.4% of a standard deviation from the mean, respectively—compared to candidates in set A (both $ps < .01$). This contrasts with direct discrimination: comparing sets A and B reveals a small and insignificant gap in both hiring likelihood and wages—in fact, the latter actually directionally $favors$ the female candidates—which replicates recent research using audit and correspondence studies to look at direct gender discrimination (Schaerer et al. 2023).

The comparison of sets B and C identifies the systemic discrimination component, as it compares the materials that female candidates could have had in the absence of language-based direct discrimination to the materials of female candidates who have these disparities in their recommendation letters. Specifically, it isolates the impact of the disparity in language stemming from direct discrimination at Node 1 on actions at the focal hiring node n^*. Systemic discrimination was large and comprised the vast majority of total discrimination: the differences in hiring likelihood and prospective wages between sets B and C corresponded to roughly 26.6% and 42.7% of a standard deviation from the mean, respectively. These results suggest that discrimination baked into non-group signals, such as the way that women



(a) Hiring Likelihood

(b) Prospective Wage

Figure 3. Experimental Iterated Audit

are described for prospective jobs, may be responsible for explaining at least some of the gender disparities observed in the labor market.

To highlight the utility the IA in this setting, we consider alternative analyses one may run to capture the impact of language disparities on evaluations. Prior work highlights differences on language associated with independence, ability, leadership, and being a standout between male and female recommendation letters (Schmader et al. 2007). One can therefore explore the impact of these differences on evaluations by regressing hiring likelihood and prospective wages on differences on these dimensions in the recommendation letters. To do so, we use the methods from Kaplan, Palitsky, Arconada Alvarez, Pozzo, Greenleaf, Atkinson, and Lam (2024) to create indices across the respective language dimensions for each of the three sets of materials used in the study. Table 2 reports regression results on total discrimination (comparing sets A and C) as a function of language attributes and the gender of the applicant. We find that pre-specified language attributes explain little of the variation in the dependent variables. Controlling for them does not mitigate almost any of the total discrimination in our study. This highlights the difficulty of examining the impact of a multi-dimensional signals such as text data, as it is not at all clear what attributes of language, including non-obvious interactions, may impact evaluations. Without the experimental IA design, a researcher can potentially fail to identify the attribute combinations that matter for evaluations and conclude that language disparities will not contribute discrimination at the focal node.

7 Conclusion

Large literatures, mostly from outside of economics, emphasize the importance of systemic factors in driving group-based disparities; yet economic analyses largely focus on direct discrimination as a function of of group identity itself. We bridge this gap by developing new

Table 2. Independent Impact of Language Constructs

	(1) Hiring Likelihood	(2) Prospective Wage
Female	-0.50**	-12.80***
	(0.23)	(3.11)
Ability	-33.30	853.27
	(44.53)	(787.32)
Leadership	22.27	19.02
	(30.46)	(428.51)
Standout	4.47	-1266.07
	(54.30)	(1083.74)
Agentic	0.02	59.88
	(24.33)	(460.43)
Constant	8.20	192.66*
	(6.57)	(113.21)
Observations	441	441
R-squared	0.024	0.030

Standard errors in parentheses, clustered at the Hiring Manager level

*** p¡0.01, ** p¡0.05, * p¡0.1

theoretical and empirical tools to study systemic discrimination. Our general theoretical framework nests the canonical notion of direct discrimination with broader notions, and formalizes the importance of researchers taking one (or several) explicit stances on individual qualification for a given action. Our empirical decomposition of total discrimination into direct and systemic components further motivates the development of new econometric tools that identify these components in experimental and observational data. Our empirical applications, including the novel constructive and experimental IA designs—show how conventional methods of studying direct discrimination can miss total discrimination and important heterogeneity in practice.

By formalizing the differences and possible interactions between direct and systemic discrimination, our framework can be useful for interpreting and predicting the effects of policies aimed at reducing disparities. Consider the case of racial or gender quotas. In standard models of taste-based or statistical discrimination, such policies would have a temporary effect on disparities: evaluators' decisions would revert back and the disparity would re-emerge when a quota is lifted. However, if the initial disparity was due to technological systemic discrimination, e.g., in access to skill development, then quotas may reduce the disparity in the skill distribution as they create an incentive to develop female players. De Sousa and Niederle (2022) show that the introduction of a team quota for the minimum number of female chess players improved the performance of female chess players across the country (but not outside the country), presumably, as the authors note, because this created an incentive to invest in the skill of female chess players.

New analytic tools may also broaden the set of appropriate policy responses to observed disparities. Systemic discrimination can lead to illegal disparate impact in some settings, as in the landmark *Griggs v. Duke Power Co. (1971)* finding. The development of robust econometric methods for measuring systemic and total discrimination, perhaps across different qualification measures, can be a powerful complement to existing regulatory tools in such settings.[31] Robust economic models of systemic discrimination can aide the interpretation of these methods, by enriching policymakers' understanding of interactions over time or across different domains. Such theoretical and empirical advancements can improve policy making in labor markets, housing, criminal justice, education, healthcare, and other areas.

References

Aaronson, D., D. Hartley, and B. Mazumder (2021): "The Effects of the 1930s HOLC 'Redlining' Maps," *American Economic Journal: Economic Policy*, 13, 355–92.

Agan, A. and S. Starr (2017): "The Effect of Criminal Records on Access to Employment," *American Economic Review P&P*, 107, 560–64.

Agan, A. Y., B. Cowgill, and L. K. Gee (2021): "Salary history and employer demand: Evidence from a two-sided audit," Tech. rep., National Bureau of Economic Research.

Aigner, D. J. and G. G. Cain (1977): "Statistical Theories of Discrimination in Labor Markets," *Industrial and Labor Relations Review*, 30, 175–187.

Allard, S. W. and M. L. Small (2013): "Reconsidering the Urban Disadvantaged: The Role of Systems, Institutions, and Organizations," *Annals of the American Academy of Political and Social Science*, 647, 6–20.

Althoff, L. and H. Reichardt (2022): "Jim Crow and Black economic progress after slavery," *Manuscript, Princeton University.*

Altonji, J. G. and R. M. Blank (1999): "Race and gender in the labor market," *Handbook of labor economics*, 3, 3143–3259.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner (2016): "Machine Bias," *ProPublica Report.*

Arnold, D., W. Dobbie, and P. Hull (2021): "Measuring Racial Discrimination in Algorithms," *AEA Papers and Proceedings*, 111, 49–54.

——— (2022): "Measuring Racial Discrimination in Bail Decisions," *American Economic Review.*

[31] For example, the U.S. Equal Employment Opportunity Commission (EEOC) launched nearly 600 investigations into systemic discrimination in 2020. Many employment practices EEOC flags for possible systemic are indirect (such as word-of-mouth recruitment practices), and would thus not be picked up by a conventional correspondence or audit study (see https://www.eeoc.gov/systemic-enforcement-eeoc).

Arnold, D., W. Dobbie, and C. S. Yang (2018): "Racial Bias in Bail Decisions," *Quarterly Journal of Economics*, 133, 1885–1932.

Arrow, K. J. (1973): "The Theory of Discrimination," in *Discrimination in Labor Markets*, ed. by O. Ashenfelter and A. Rees, Princeton, NJ: Princeton University Press.

Baron, E. J., J. J. Doyle Jr, N. Emanuel, P. Hull, and J. P. Ryan (2023): "Racial Discrimination in Child Protection," Tech. rep., National Bureau of Economic Research.

Barron, K., R. Ditlmann, S. Gehrig, and S. Schweighofer-Kodritsch (2020): "Explicit and Implicit Belief-Based Gender Discrimination: a Hiring Experiment," *WZB Discussion Paper*.

Bartik, A. W. and S. T. Nelson (2016): *Credit reports as resumes: The incidence of pre-employment credit screening*, Massachusetts Institute of Technology, Department of Economics.

Becker, G. S. (1957): *The Economics of Discrimination*, University of Chicago Press.

Berk, R., H. Heidari, S. Jabbari, M. Kearns, and A. Roth (2018): "Fairness in Criminal Justice Risk Assessments: The State of the Art," *Sociological Methods & Research*, 50, 1–42.

Bertrand, M. and E. Duflo (2016): *Field Experiments on Discrimination*, Elsevier.

Bertrand, M. and S. Mullainathan (2004a): "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review*, 94, 991–1013.

——— (2004b): "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination," *American economic review*, 94, 991–1013.

Bierer, B. E., L. G. Meloney, H. R. Ahmed, and S. A. White (2022): "Advancing the inclusion of underrepresented women in clinical research," *Cell Reports Medicine*, 3.

Blinder, A. S. (1973): "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, 8, 436–455.

Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2022): "Inaccurate Statistical Discrimination: An Identification Problem," *Review of Economics and Statistics*.

Bohren, J. A., A. Imas, and M. Rosenberg (2019): "The Dynamics of Discrimination: Theory and Evidence," *American Economic Review*, 109, 3395–3436.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016): "Stereotypes," *The Quarterly Journal of Economics*, 131, 1753–1794.

——— (2019): "Beliefs About Gender," *American Economic Review*, 109, 739–73.

Brekoulakis, S. (2013): "Systemic Bias and the Institution of International Arbitration: A New Approach to Arbitral Decision-Making," *Journal of International Dispute Settlement*, 4, 553–585.

Buchmann, N., C. Meyer, and C. D. Sullivan (2023): "Paternalistic Discrimination," Working Paper.

Buolamwini, J. (2022): "Facing the Coded Gaze with Evocative Audits and Algorithmic Audits," Ph.D. thesis, Massachusetts Institute of Technology.

Cain, G. G. (1986): *The Economic Analysis of Labor Market Discrimination: A Survey*, Elsevier.

Ceci, S. J. and W. M. Williams (2011): "Understanding current causes of women's underrepresentation in science," *Proceedings of the National academy of sciences*, 108, 3157–3162.

Coate, S. and G. C. Loury (1993): "Will Affirmative-Action Polices Eliminate Negative Stereotypes?" *American Economic Review*, 83, 1220–1240.

Coffman, K. B., C. L. Exley, and M. Niederle (2021): "The Role of Beliefs in Driving Gender Discrimination," *Management Science*, 67, 3551–3569.

Conway, J., R. Mill, and L. C. Stein (2023): "The Gendered Impacts of Perceived Skin Tone: Evidence from African American Siblings in 1870–1940," .

Cook, C., R. Diamond, J. V. Hall, J. A. List, and P. Oyer (2021): "The gender earnings gap in the gig economy: Evidence from over a million rideshare drivers," *The Review of Economic Studies*, 88, 2210–2238.

Cook, L. (2014): "Violence and Economic Growth: Evidence from African American Patents, 1870-1940," *Journal of Economic Growth*, 19, 221–257.

Cornell, B. and I. Welch (1996): "Culture, Information, and Screening Discrimination," *Journal of Political Economy*, 104, 542–571.

Crawley, L. M., D. K. Ahn, and M. A. Winkleby (2008): "Perceived medical discrimination and cancer screening behaviors of racial and ethnic minority adults," *Cancer Epidemiology Biomarkers & Prevention*, 17, 1937–1944.

Darity, W. (2005): "Stratification Economics: the Role of Intergroup Inequality," *Journal of Economics and Finance*, 29, 144–153.

Darity, W. A. and P. L. Mason (1998): "Evidence on Discrimination in Employment: Codes of Color, Codes of Gender," *Journal of Economic Perspectives*, 12, 63–90.

De Plevitz, L. (2007): "Systemic Racism: The Hidden Barrier to Educational Success for Indigenous School Students," *Australian Journal of Education*, 51, 54–71.

De Quidt, J., J. Haushofer, and C. Roth (2018): "Measuring and Bounding Experimenter Demand," *American Economic Review*, 108, 3266–3302.

De Sousa, J. and M. Niederle (2022): "Trickle-down effects of affirmative action: A case study in France," Tech. rep., National Bureau of Economic Research.

Derenoncourt, E., C. H. Kim, M. Kuhn, and M. Schularick (2022): "Wealth of two nations: The US racial wealth gap, 1860-2020," Tech. rep., National Bureau of Economic Research.

Eli, S. J., T. D. Logan, and B. Miloucheva (2023): "The Enduring Effects of Racial Discrimination on Income and Health," *Journal of Economic Literature*, 61, 924–940.

Fang, H. and A. Moro (2011): "Theories of Statistical Discrimination and Affirmative Action: A Survey," in *Handbook of Social Economics*, Elsevier, vol. 1, 133–200.

Feagin, J. (2013): *Systemic Racism: A Theory of Oppression*, Routledge.

Feagin, J. R. and C. B. Feagin (1978): *Discrimination American Style: Institutional Racism and Sexism*, Prentice Hall.

Fiske, S. T. (1998): "Stereotyping, Prejudice, and Discrimination," *The Handbook of Social Psychology*, 2, 357–411.

Gallen, Y. and M. Wasserman (2021): "Informed Choices: Gender Gaps in Career Advice," *CEPR Discussion Paper No. DP15728.*

Gawai, V. P. and J. D. Foltz (2023): "Discrimination in Science: Salaries of Foreign and US Born Land-Grant University Scientists," .

Gebru, T. (2020): "Race and gender," *The Oxford handbook of ethics of aI*, 251–269.

Glover, D., A. Pallais, and W. Pariente (2017): "Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores," *The Quarterly Journal of Economics*, 132, 1219–1260.

Grau, N. and D. Vergara (2021): "An Observational Implementation of the Outcome Test with an Application to Ethnic Prejudice in Pretrial Detentions," *Working Paper.*

Gynter, P. (2003): "On the Doctrine of Systemic Discrimination and its Usability in the Field of Education," *International Journal on Minority and Group Rights*, 10, 45–54.

Hardt, M., E. Price, and N. Srebro (2016): "Equality of Opportunity in Supervised Learning," *Proceedings of the 30th Conference on Neural Information Processing Systems*, 3323–3331.

Harrington, E. and H. Shaffer (2023): "Brokers of Bias in the Criminal System: Do Prosecutors Compound or Attenuate Disparities Inherited from Arrest?" .

Hill, R. B. (1988): *Structural Discrimination: The Unintended Consequences of Institutional Processes.*, Wesleyan University Press.

Hübert, R. and A. T. Little (2020): "A Behavioral Theory of Discrimination in Policing," *Working Paper.*

Hull, P. (2021): "What Marginal Outcome Tests Can Tell Us About Racially Biased Decision-Making," *NBER Working Paper No. 28503.*

Hurst, E., Y. Rubinstein, and K. Shimizu (2021): "Task-based discrimination," Tech. rep., National Bureau of Economic Research.

Jerant, A. F., J. J. Fenton, and P. Franks (2008): "Determinants of racial/ethnic colorectal cancer screening disparities," *Archives of internal medicine*, 168, 1317–1324.

Jolly-Ryan, J. (1998): "Chipping Away at Discrimination at the Country Club," *Pepper-*

dine Law Review, 25, 2.

Kaplan, D. M., R. Palitsky, S. J. Arconada Alvarez, N. S. Pozzo, M. N. Greenleaf, C. A. Atkinson, and W. A. Lam (2024): "What's in a Name? Experimental Evidence of Gender Bias in Recommendation Letters Generated by ChatGPT," *Journal of Medical Internet Research*, 26, e51837.

Kasy, M. and R. Abebe (2021): "Fairness, equality, and power in algorithmic decision-making," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 576–586.

Kessler, J. B., C. Low, and C. D. Sullivan (2019): "Incentivized resume rating: Eliciting employer preferences without deception," *American Economic Review*, 109, 3713–44.

Kitagawa, E. M. (1955): "Components of a Difference Between Two Rates," *Journal of the American Statistical Association*, 50, 1168–1194.

Kline, P. M., E. K. Rose, and C. R. Walters (2021): "Systemic Discrimination among Large U.S. Employers," *NBER Working Paper No. 29053.*

——— (2024): "A discrimination report card," Tech. rep., National Bureau of Economic Research.

Knowles, J., N. Persico, and P. Todd (2001): "Racial Bias in Motor Vehicle Searches: Theory and Evidence," *Journal of Political Economy*, 109, 203–229.

Kübler, D., J. Schmid, and R. Stüber (2018): "Gender discrimination in hiring across occupations: a nationally-representative vignette study," *Labour Economics*, 55, 215–229.

Kupfer, S. S., S. McCaffrey, and K. E. Kim (2006): "Racial and gender disparities in hereditary colorectal cancer risk assessment: the role of family history." *Journal of Cancer Education*, 21.

Lahey, J. N. and D. R. Oxley (2021): "Discrimination at the Intersection of Age, Race, and Gender: Evidence from an Eye-Tracking Experiment," *Journal of Policy Analysis and Management*, 40, 1083–1119.

List, J. A. (2004): "The nature and extent of discrimination in the marketplace: Evidence from the field," *The Quarterly Journal of Economics*, 119, 49–89.

Lodermeier, A. (2023): "Racial Discrimination in Eviction Filing," .

Mayhew, L. H. (1968): *Law and Equal Opportunity*, Harvard University Press.

Mengel, F., J. Sauermann, and U. Zölitz (2019): "Gender Bias in Teaching Evaluations," *Journal of the European Economic Association*, 17, 535–566.

Mocanu, T. (2022): "Designing Gender Equity: Evidence from Hiring Practices and Committees," .

National Archive (2007): "Access to Archival Databases (AAD)," .

Neal, D. A. and W. R. Johnson (1996): "The Role of Premarket Factors in Black-White

Wage Differences," *Journal of Political Economy*, 104, 869–895.

Oaxaca, R. (1973): "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, 14, 693–709.

Pager, D., B. Bonikowski, and B. Western (2009): "Discrimination in a Low-Wage Labor Market: A Field Experiment," *American Sociological Review*, 74, 777–799.

Pager, D. and H. Shepherd (2008): "The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets," *Annuual Review of Sociology*, 34, 181–209.

Phelps, E. S. (1972): "The Statistical Theory of Racism and Sexism," *American Economic Review*, 62, 659–661.

Pierson, E., C. Simoiu, J. Overgoor, S. Corbett-Davies, D. Jenson, A. Shoemaker, V. Ramachandran, P. Barghouty, C. Phillips, R. Shroff, et al. (2020): "A Large-Scale Analysis of Racial Disparities in Police Stops Across the United States," *Nature Human Behaviour*, 4, 736–745.

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

Pinkston, J. C. (2003): "Screening discrimination and the determinants of wages," *Labour Economics*, 10, 643–658.

Powell, J. A. (2007): "Structural Racism: Building Upon the Insights of John Calmore," *North Carolina Law Review*, 86, 791.

Rambachan, A. and J. Roth (2020): "Bias In, Bias Out? Evaluating the Folk Wisdom," *1st Symposium on the Foundations of Responsible Computing (FORC 2020)*, 156, 6:1–6:15.

Rose, E. K. (2022): "A Constructivist Perspective on Empirical Discrimination Research," *Working Paper.*

Rothstein, R. (2017): *The Color of Law: A Forgotten History of How Our Government Segregated America*, Liveright Publishing.

Sarsons, H. (2019): "Interpreting Signals in the Labor Market: Evidence from Medical Referrals," *Working Paper.*

Schaerer, M., C. Du Plessis, M. H. B. Nguyen, R. C. Van Aert, L. Tiokhin, D. Lakens, E. G. Clemente, T. Pfeiffer, A. Dreber, M. Johannesson, et al. (2023): "On the trajectory of discrimination: A meta-analysis and forecasting survey capturing 44 years of field experiments on gender and hiring decisions," *Organizational Behavior and Human Decision Processes*, 179, 104280.

Schmader, T., J. Whitehead, and V. H. Wysocki (2007): "A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants," *Sex roles*, 57, 509–514.

Small, M. L. and D. Pager (2020): "Sociological perspectives on racial discrimination," *Journal of Economic Perspectives*, 34, 49–67.

Trix, F. and C. Psenka (2003): “Exploring the color of glass: Letters of recommendation for female and male medical faculty,” *Discourse & Society*, 14, 191–220.

Wan, Y., G. Pu, J. Sun, A. Garimella, K.-W. Chang, and N. Peng (2023): “” kelly is a warm person, joseph is a role model”: Gender biases in llm-generated reference letters,” *arXiv preprint arXiv:2310.09219*.

Williams, J. A., T. D. Logan, and B. L. Hardy (2021): “The persistence of historical racial violence and political suppression: Implications for contemporary regional inequality,” *The ANNALS of the American Academy of Political and Social Science*, 694, 92–107.

Zafar, M. B., I. Valera, M. Gomez Rodriguez, and K. Gummadi (2017): “Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment,” *Proceedings of the 26th International Conference on World Wide Web*.

Zink, A., Z. Obermeyer, and E. Pierson (2023): “Race Corrections in Clinical Models: Examining Family History and Cancer Risk,” *medRxiv*, 2023–03.

Zivin, J. S. G. and G. Singer (2023): “Disparities in pollution capitalization rates: The role of direct and systemic discrimination,” Tech. rep., National Bureau of Economic Research.

A Related Literature

Our framework builds on a large literature studying the role of systemic forces in driving group-based disparities (e.g., Pincus 1996; Feagin 2013; Allard and Small 2013; Pager and Shepherd 2008). While exact definitions vary (Small and Pager 2020), this systems-based approach distinguishes between direct discrimination, where individuals or firms treat people differently because of group identity itself, and indirect or systemic discrimination that considers the interlocking institutions or domains through which inequities propagate (Gynter 2003). In the systems-based approach, channels for observed disparities are taken as cumulative both within and across domains; discrimination is not just a product of a single individual or institution (Powell 2007). Systemic (or "structural") discrimination can be generated by the indirect relationships between outcomes and evaluations in roughly the same period, such as when discrimination in criminal justice drives unwarranted disparities in education and labor market outcomes.[32] It is also generated over time, such as when historic "redlining" practices in lending generates persistent disparities in credit access through its differential effects on generational wealth (e.g., Aaronson, Hartley, and Mazumder (2021)). The literature sometimes refers to the former as "side-effect" discrimination and the latter as "past-in-present" discrimination (Gynter 2003; Feagin and Feagin 1978; Feagin 2013).

Importantly, the systemic perspective shifts focus from the motives and biases of a given individual or institution to policies or institutional arrangements that contribute to *de facto* discrimination, perhaps without intent. Direct discrimination, either on the part of individuals or institutions, is inherently non-neutral: it arises from the explicit differential treatment of individuals on the basis of group identity. Systemic discrimination, in contrast, can exist in policies that are facially neutral by race, gender, or other protected characteristics (Hill 1988). For example, a lending algorithm which considers a person's zip code but does not use racial information when determining loan eligibility may be race neutral in design but discriminatory in practice. Black borrowers may be more likely to live in certain zip codes than equally creditworthy white borrowers, perhaps because of prior discriminatory policies in housing, employment, or financial markets.[33]

The distinction between direct and indirect discrimination is echoed in legal theories of disparate treatment and disparate impact (e.g., Brekoulakis 2013; Gynter 2003; De Plevitz 2007; Rothstein 2017). Under the disparate impact doctrine, a policy or practice may be deemed discriminatory if it leads to disparities without substantial legitimate justification—

[32] Powell (2007) considers systemic discrimination as driving disparities within a domain, e.g., the hiring and promotion practices within a firm or industry, and structural discrimination as driving disparities through the interaction of different systems.

[33] Note that policies that are facially neutral on protected characteristics may not be neutral in intent. Mayhew (1968) argues that some organizations may have accepted Civil Rights legislation mandating "colorblind" treatment because they were aware systemic discrimination could preserve the status quo.

as in *Griggs v. Duke Power Co. (1971)*.[34] A facially neutral practice may therefore be found to be discriminatory under this doctrine even in the absence of explicit categorization or animus. This notion of discrimination contrasts with the disparate treatment doctrine, which prohibits policies or practices motivated by a discriminatory purpose. Typically, proof of discriminatory intent is required for the finding of disparate treatment.[35]

A systemic perspective is also found in the recent literature on algorithmic unfairness (e.g., Angwin et al. 2016; Hardt, Price, and Srebro 2016; Zafar, Valera, Gomez Rodriguez, and Gummadi 2017; Berk et al. 2018; Kasy and Abebe 2021; Gebru 2020; Buolamwini 2022; Arnold, Dobbie, and Hull 2021). An algorithm which does not directly use protected characteristics may nevertheless return systematically disparate outcome predictions or treatment recommendations among equally qualified individuals. The literature studies how interlocking systems of data collection, model fitting, and human-algorithm decision-making may generate such disparities.

Finally, research in the field of stratification economics proposes a systemic perspective as necessary for understanding group-based disparities because advantaged groups have an incentive to maintain them (Darity 2005; Darity and Mason 1998; De Quidt, Haushofer, and Roth 2018). Without considering the systemic interactions generating a specific outcome, as well as the incentives involved in maintaining this system, a researcher or policy maker may miss important channels through which group-based disparities persist.

Our work also adds to the long literature on direct discrimination in economics, which is typically modeled as a causal effect of group membership on treatment.[36] Theoretical sources of direct discrimination include individual preferences or beliefs. In the canonical framework of taste-based discrimination, differential treatment emerges because individuals derive disutility from interacting with or providing services to members of a particular group (Becker 1957). In models of belief-based discrimination, differential treatment emerges because a decision-relevant statistic (such as labor market productivity) is unobserved, and there are group-based differences in beliefs about its distribution (Phelps 1972; Arrow 1973; Aigner and Cain 1977). While belief differences have traditionally been assumed to stem from true differences in the distributions, a recent literature has considered the role of inaccurate beliefs in driving direct discrimination (Bohren et al. 2022; Barron, Ditlmann, Gehrig, and Schweighofer-Kodritsch 2020; Hübert and Little 2020). These differences may stem from a lack of information or biased stereotypes (Bordalo, Coffman, Gennaioli, and Shleifer 2016; Coffman, Exley, and Niederle 2021; Bordalo et al. 2019; Fiske 1998), which again lead to

[34] See also *Dothard v. Rawlinson (1977)* and *Cocks v. Queensland (1994)*

[35] See, e.g., *Washington v. Davis (1976)* and *McClesky v. Kemp (1987)*.

[36] Notable exceptions to the typical focus on direct discrimination in economics include Neal and Johnson (1996), Glover, Pallais, and Pariente (2017), List (2004), Cook (2014), Hurst et al. (2021), and Sarsons (2019). In Section 3.4 we discuss how the model of Coate and Loury (1993) captures a specific source of systemic discrimination in our framework.

causal effects of a protected characteristic on evaluations and decision-making.

A rich empirical literature in economics has largely followed this theoretical tradition. Research using both experimental and observational data has attempted to identify the causal effect of group identity on treatment, holding other observables constant (e.g., Bertrand and Mullainathan 2004a; Fang and Moro 2011; Bertrand and Duflo 2016). In the widely-used correspondence study method, evaluators (e.g., hiring managers) are presented with information about individuals (e.g., applicants for a job), which consists of the individual's group identity and other signals of their qualifications (e.g., education level). Since everything but group identity—or a signal of this identity—is held constant in the experimental design, any differential treatment can be directly attributed to the causal effect of this variable. Recent advances in this methodology have been used to examine the dynamics of discrimination (Bohren et al. 2019) and the heterogeneity in discrimination across institutions (Kline et al. 2021).[37] A parallel empirical literature has developed tools to distinguish different economic theories of discrimination. Recent advances involve outcome tests of racial bias, in both observational (Knowles, Persico, and Todd 2001; Grau and Vergara 2021) and quasi-experimental data (Arnold, Dobbie, and Yang 2018; Hull 2021).

As also noted in Small and Pager (2020), the systemic perspective suggests that standard tools for measuring direct discrimination miss an important component. Efforts to model and measure causation at any particular juncture and within a specific domain can substantially understate the cumulative impact of discrimination across domains or time. We contribute to the economics literature by expanding the tools for studying such forms of discrimination. Additionally, our framework offers new interpretations for previously documented group-based disparities. For example, evidence for a reversal of direct discrimination over time—such as the ones documented in Bohren et al. (2019) and Mengel, Sauermann, and Zölitz (2019)—may not imply that total discrimination has been mitigated or reversed. If, as argued, biased evaluators drive initial discrimination in the pipeline, the group that ends up being favored may still face substantial total discrimination when conditioning on underlying qualifications.[38]

A small but growing literature in economics has examined the impact of previous direct discrimination on subsequent disparities. Cook (2014) and Williams et al. (2021) study the long-run effects of racial violence on innovation and regional inequality, respectively. Eli

[37] While Kline et al. (2021) refer to their study as estimating "systemic discrimination," this classification is not consistent with the large social science literature on systemic discrimination outlined above. Their correspondence study is designed to measure direct discrimination, formalized as the causal effects of protected characteristics in a hiring decision. We view this work as more accurately studying institutional direct discrimination.

[38] The systemic perspective also highlights the lasting impact of initial stereotypes (Bordalo et al. 2016, 2019). Even if signals become more precise and direct discrimination decreases, total discrimination can persist through systemic channels.

et al. (2023) and Derenoncourt et al. (2022) review and examine the impact of historical discriminatory practices on the evolution of the racial wealth gap.

A series of papers have built directly on our definitions and framework to measure and classify direct, systemic, and total discrimination. Althoff and Reichardt (2022) measure the systemic components of disparities that stem from racially oppressive institutions—slavery and Jim Crow laws. Baron et al. (2023) examine discrimination in foster care through the investigator-screener relationship, finding that systemic discrimination generated by screeners accounts for a substantial proportion of the resulting total discrimination. Zivin and Singer (2023) study racial differences in home values as a function of pollution exposure, concluding that 75% of the disparity was driven by systemic discrimination in complementary amenities. Lodermeier (2023) applies our framework to the study of eviction rates, finding that the substantial racial disparity is likely caused by direct rather than systemic discrimination. Gawai and Foltz (2023) look at the impact of country of birth on income in academia and find significant total discrimination. They identify two-thirds of that disparity to be driven by systemic discrimination. Finally, Buchmann et al. (2023) study a form of anticipated systemic discrimination where employers are less likely to hire women due to gender-based disparities in safety outside of the job, which they term *paternalistic discrimination*. They find that eliminating this type of discrimination would reduce the gender employment gap by 24% and increase female wages by 21% in their setting.

B Screening Discrimination

We proceed to illustrate screening discrimination empirically in an online labor market, using a setup similar to the one in **??**.

Similar to the case of statistical direct discrimination (e.g., Fang and Moro 2011), differential signal precision can be heterogeneous across qualification level. Consider, for example, a hiring decision in which the signal is equal to productivity plus mean-zero noise. A noisier signal hurts high productivity workers, as it leads to a higher chance of generating a signal below the hiring threshold, but can benefit low productivity workers by leading to a higher chance of a generating a signal above the hiring threshold. In contrast, in a medical diagnostic decision, all patients benefit from a more accurate signal when it leads to more accurate diagnoses regardless of health status.

Experimental Setup

This experiment used the same group of Workers as in **??**. A new group of 199 Recruiters were shown the task-A performance of two Workers, along with the Workers' gender, and asked to select which Worker they would prefer to hire. Recruiters were then paid 1 USD for each question the hired Worker answered correctly on task B, above 5. The Recruiter's action rule is thus $A_i^R \in \{0, 1\}$.

A new group of 501 Hiring Managers saw one Worker's profile after their evaluation by a Recruiter, along with the Worker's gender. They were shown information on the Worker's task-A performance only if the Recruiter had chosen to hire them; otherwise they saw no performance information. Therefore, $\mathcal{S}^H = \{\emptyset, 2, 3, 4, 5, 6\}$. Hiring Managers then made a binary decision of whether or not to hire the Worker. If the Worker was hired, the Hiring Manager received a bonus corresponding to their task-B performance; otherwise, the Hiring Manager received 4 dollars with certainty.

Formally, each Hiring Manager j observed a signal S_i^H corresponding to Worker i's task-A performance if the Worker was hired by the recruiter ($A_i^R = 1$). If the Worker was not hired ($A_i^R = 0$), the Hiring Manager observed no signal ($S_i^H = \emptyset$). Recruiter actions thus affected the *informativeness* of Hiring Manager signals—whether or not she saw an objective signal of productivity. This setting was designed to emulate the process by which managers can obtain more accurate performance signals depending on whether potential Workers had access to prior opportunities to "prove themselves" (e.g., internships). The Manager's action $A_i^H \in \{0, 1\}$ corresponds to her hiring the Worker.

Results

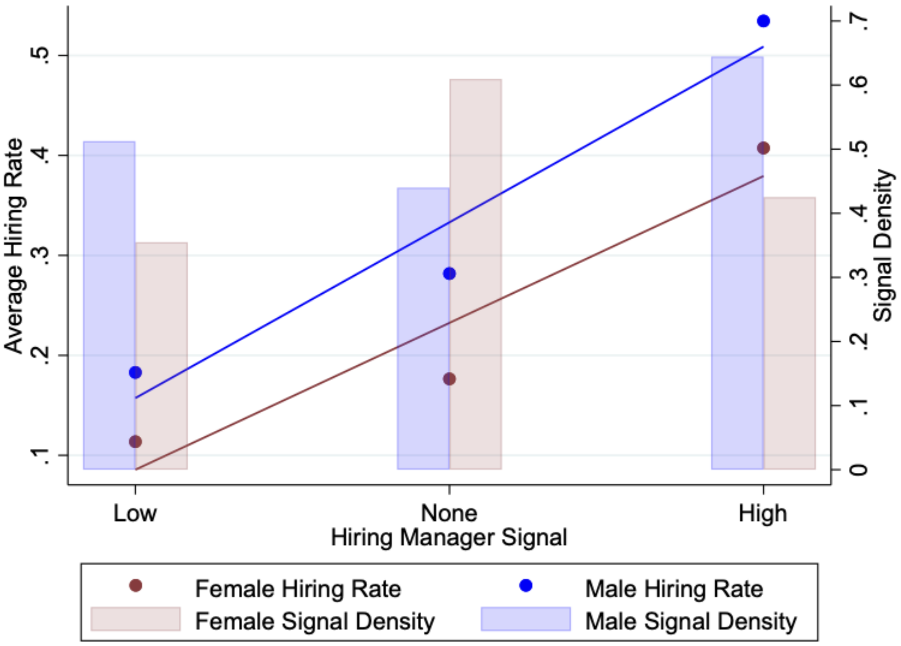
As before, we measure systemic and total discrimination with respect to task-A performance, $Y_i^0 = S_i^R$, with $\mathcal{Y}^0 = \{2, 3, 4, 5, 6\}$. Since this qualification measure coincides with the Recruiter signal, any discrimination in the Recruiter stage is direct. Discrimination in the Hiring Manager stage can again be direct or systemic. Per **??**, we expect the differences in signal informativeness to lead to heterogeneity in systemic discrimination by qualification.

Recruiters: Recruiters directly discriminated against female Workers. The hiring rate for male Workers was 28 percentage points higher than for female Workers ($p < 0.01$), who were hired at a rate of 36%.[39] Given the lack of gender-based performance differences, as reported in **??**, this disparity in hiring rates is not consistent with accurate statistical discrimination. Therefore, Recruiter direct discrimination again stems from either biased preferences or beliefs.

Hiring Managers: Hiring Managers discriminated against female Workers. On average, male Workers were hired at a 9 percentage point higher rate than female Workers ($p = 0.02$), who were hired at a rate of 0.22. However, this average effect masks important heterogeneity. Among Workers with low (below-median) qualification levels, male Workers were hired at an insignificant 4 percentage point higher rate ($p = 0.43$).[40] Among Workers with high (above-median) qualification levels, male Workers were hired at a significant 23 percentage point higher rate ($p < 0.01$).

[39] Standard errors are clustered at the individual level.

[40] The median task-A performance was 4.



Notes: This figure plots average Hiring Manager hiring rates (left y-axis) and signal probabilities (right y-axis) by productivity signal for female and male workers, where high versus low signal corresponded to either above and equal to or below the median (3), respectively. Gender differences in the hiring rates for a given signal illustrates direct discrimination, while gender differences in the signal probability illustrates the source of systemic discrimination.

FIGURE 4. Screening: Hiring Manager Hiring Rate and Signals

Figure 4 illustrates the reason for this heterogeneity in total discrimination. Similar to **??**, the scatter plot shows the average Hiring Manager actions conditional on the signal (or lack thereof) and the Worker's gender. As before, the lines of best fit show a positive relationship between the signal and the probability of getting hired for both groups: Hiring Managers were more likely to hire a Worker after seeing a high signal than a low signal, with the hiring rate for no signal laying in between. Conditional hiring rates are shifted upward for male Workers, illustrating direct discrimination. Importantly, however, the distribution of signals seen by Managers also differs by gender: direct discrimination by Recruiters made Managers more likely to see both low and high signals from male Workers than female Workers, with female Workers being much more likely to have an uninformative signal. Given the upward-sloping lines, female Workers with high qualification levels were likely to be hurt by systemic discrimination, while female Workers with low qualification levels were likely to be helped by it.

We quantify total, direct, and systemic discrimination in Hiring Manager actions using the decompositions in Section 3.5. We estimate Hiring Manager total discrimination $\Delta(y^0)$ by comparing male and female hiring rates based on task-A performance. We then estimate the Hiring Manager average direct discrimination $\overline{\tau}(w, y^0)$ faced by male Workers with a

Table 3. Screening: Discrimination Decomposition

	High Qualification (1)	Low Qualification (2)	Difference (3)
Total	0.24***	0.03	0.21***
	(0.06)	(0.04)	(0.07)
Average Direct	0.15***	0.07**	0.08
	(0.05)	(0.04)	(0.05)
Systemic	0.09**	-0.04	0.13**
	(0.04)	(0.03)	(0.06)
# Observations	501	501	501

Notes: This table reports estimates of each measure of discrimination in Equation (6) for Hiring Manager hiring rates, averaged by an equal-weighted distribution of task-A scores for male and female Workers in the given qualification bin, where High corresponds to above or equal to the median (3) and Low corresponds to below the median. Total discrimination is measured by the average difference in hiring rates among male versus female Workers with a given task-A score. The sample includes 501 Hiring Managers, each evaluating one Worker. Robust standard errors, obtained from a weighted bootstrap, are reported in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

given task-A performance by averaging gender disparities across each Hiring Manager signal realization according to the distribution each task-A performance induces over this signal. Subtracting this estimate of from the estimate of total discrimination yields an estimate of the measure of systemic discrimination.[41] We average these measures over the distribution of task-A performance as before, separately for Workers with low (below-median) and high (above-median) qualification levels.

Table 3 confirms the heterogeneity in systemic discrimination faced by women with different qualification levels. For highly qualified women, total discrimination is estimated as a significant 0.24. Our decomposition shows this is driven by a combination of significant direct (0.15) and systemic discrimination (0.09). In contrast, total discrimination among workers with a low qualification is small and insignificant (0.03), despite significant direct discrimination. The reason is a small degree of negative systemic discrimination among less qualified Workers (-0.04). Consistent with the model in **??**, the gap in systemic discrimination across qualification levels is significant ($p = 0.04$).

[41] Here we use the "average" decomposition, Equation (5). The other decompositions give similar results.

Experimental Details

B.1 Constructive IA

B.1.1 Scraping job listings

For the IA Study

We scraped job listings from corporate career websites of five major automotive firms: AutoZone, O'Reilly, Advance Auto, AutoNation, and Napa. For each company, we scraped open entry-level job listings for full-time sales position with minimal requirements (high school and no experience). In general, for each career website, we filtered on full-time status, job category, and posted time (within the past two weeks) when available. After the results were filtered, our scraping script went through each page and recorded each job listing's title and URL of its detailed job listing page.

The job requirements were checked as the script recorded the job titles and links. After each page, a list of unique job titles and their URLs (randomly selected if there were more than one listing under the same title) was created, and each URL within this list was visited to check for job requirements. The script evaluated whether the job listing qualified as entry-level by searching for keywords: first, it located all sentences/clauses containing keywords such as "Required", "High School", or "Ability"; second, it went through these sentences/clauses searching for keywords such as (variations of) "high school", "degree", and "experience". Only jobs requiring just a high school degree/GED or with no requirements would count as entry-level. We counted "requiring automotive knowledge" and "experience/degree preferred" as entry-level too, since all resumes mentioned automotive experience in the former case, and the requirement was not strict in the latter case. A separate list kept all the job titles whose requirements qualified as entry-level. After each page, all job listings whose titles were in the qualified list would then be included, and job titles that were not in the list (either new jobs or disqualifying jobs) would be checked. Given the disparities in language and format these companies use in their job listings, the keywords and filters were modified on a company-by-company basis. A research assistant also monitored the scraping process and manually checked job titles. Qualified jobs were matched with their store locations either from the job listings when the script recorded them in the first place (were the addresses present), or matched by searching the store identifier through the company store locators. Eventually, we randomly selected one job listing from each store for stores with multiple job listings such that all store would only see one set of four resumes.

For Entry-level Jobs in General

After implementing the audit experiment, we scraped job listings of all available entry-level jobs from the same automotive companies in the cities where we applied to. The script for this part of scraping was very similar to the script we used for recording jobs for the audit experiment, with the exceptions that we did not restrict job categories but instead restricted store locations. For example, we applied to stores in Houston, TX, and here we scraped all entry-level jobs (not just sales) in stores in Houston, TX, such as sales and drivers. For companies allowing us to specify a city, we searched for jobs labelled as in the specific city; for companies allowing us to search for jobs within certain radius to a location, we chose jobs within 5 miles of the city (the smallest radius). A research assistant similarly monitored the process.

B.1.2 Creating Resumes

Address

Using the store locations we recorded from the job listings, we assigned a residential address to each applicant. We used the National Address Database (NAD), restricting the sample to only contain addresses whose type is residential. For each job, we reverse geocoded the store location to latitude and longitude, and created a subsample of the residential addresses whose latitude and longitude are both within ± 0.29 degree of the job location latitude, and ± 0.37 degree of the job location longitude. Such degrees approximately correspond to 20 miles in distance. Further, we assigned a ordinal variable to the addresses, which takes on value 0 if the address is in the same city as the job location, 1 if the address is in the same state but not the same city, and 2 if in different states. We first ordered this subsample by the ordinal variable, such that addresses in the same city and state were ranked the highest. We randomly selected four distinct addresses from the top 200 addresses; if there were fewer than four addresses in this subsample, we use a larger subsample by changing the ±20 miles to ±30 miles and randomly selected four distinct addresses from the top 200 addresses. In practice we rarely have 200 rows to randomize from given how few residences are close to the stores. Further, since this ranking was based on city and state rather than distance, our method simply prioritized addresses in the same city/state within certain distance (disciplined by latitude and longitude). After this step, there existed jobs in locations where there were not enough residences, and we excluded these jobs to avoid very small towns.

Prior Employment

Two of each set of resumes had prior related job experiences. We assigned previous employers to such resumes using the 2022 InfoGroup business data. We filtered the data by primary SIC code to include only companies in the automotive industry. We also relabelled companies such as *Carquest* who experienced mergers with companies in the scraped jobs to avoid applying with job experience from the same companies. The next steps were identical to the steps in assigning nearby residences: we first looked at companies within ± 20 miles of the latitude and longitude of the job locations, assigned an ordinal variable based on city/state, and randomly selected four employers from the top 200 nearby companies ranked by the ordinal variavle (that were not the same as the job listing company). If there were not enough companies within ± 20 miles, the same process was repeated for the subsample within ± 30 miles. Similarly, we rarely encountered jobs with over 200 other employers nearby. Jobs with fewer than four other employers nearby were also excluded.

The start and end dates of prior employments were randomized. For each applicant with prior experience, the length of prior experience (in years) was randomly drawn from uniform $[0.5, 2]$ and rounded to the first decimal point. Correspondingly, we counted the days between the day we made the resumes and the applicant's high school graduate day, and subtracted the length of prior experience (in days) to get the applicant's unemployment days. These unemployment days were then randomly split into before and after the prior employment. The proportion of the unemployed days before the prior employment was randomly selected from uniform $[0, 1]$. The employment start date was then the high school graduate date plus these unemployed days before prior employment, and the employment end date was the start date plus the length of the prior employment (in days).

High School

All of our applicants were high school graduates with automotive-related workshop experience in high school. Only the names of such workshops were mentioned as part of the high school experiences without any description: "Automotive Technology Essentials", "Automotive Diagnostics and Service", "Car Care and Repair", and "Automotive Technology Workshop". These names were generated by ChatGPT and manually edited. All four applications to the same job used different names for the workshop to avoid suspicion. To assign each resume a high school, we used the National Center for Education Statistics (NCES) public school data. We further filtered this public school data to contain schools offering the 12th grade as the highest grade, whose type is either regular or technical school, and whose school names are not suggestive of online or art schools. For each job, we constructed a subsample of high schools whose zip code is within ± 100 of the job location zip code. We used ± 100 to locate high schools reasonably close to the store locations. For example, consider a zip code of 60637, the range would be $[60537, 60737]$. Similar to the residential addresses, we then assigned an ordinal variable to the high school depending on whether the high school is in the same city and/or state. Ranking the high schools using this ordinal variable, we randomly selected four distinct high schools from the top 200 schools. Likewise, we rarely ran into job locations with over 200 schools nearby to randomly select from, and we prioritized schools in the same city/state. This step also ruled out some job listings with fewer than four high schools nearby.

The duration of high school was randomized. For each applicant, we first randomly selected an integer from uniform $[17, 19]$ as the age when this applicant graduated from high school. We added this age to the applicant's birth year, and added a randomly selected number of days from uniform $[170, 220]$ to the first day of that year as the end date to ensure the end date was in June, July, or August. Similarly, for high school start date, we subtracted 3 from the graduation year, and added to the first day of that year a randomly selected number of days from $[210, 250]$ to ensure the start date was in August or September.

Name

To create the treatment by race, we used the race-salient names and surnames from Kline et al. (2024). We created a dataset with all combinations of the first names and surnames for each race, and randomly selected distinct names from these names for each company. The number of randomly selected names depended on the number of job listings each company had, such that no company would see the same name more than once. This is because all the companies we applied to used a central online application system and we were unsure whether duplicated names would cause potential issues.

Email

All resumes had the applicants' emails listed. To create corresponding emails for the fictitious applicants, we purchased domains and hosted emails ourselves. All applicants' emails ended with "@mailprofessional.live" and "@voyagemail.pro". The username for each applicant was randomly chosen from four potential four formats: fist-name-surname and a random integer, surname-first-name and a random integer, first-name-initial-surname and a random integer, and surname-first-name-initial and a random integer. All the integers were smaller than 10,000. If a username was too long, it would be replaced with a shorter version by randomly selecting from the last two formats only, and by randomly selecting a smaller number. After these email addresses were constructed, we first registered emails manually and another scraping script later registered all these emails on the hosting platform. A research assistant manually checked the emails. In some resume formats, the emails were too long and they were broken across two lines with a hyphen in between. These email addresses, with hyphens added, were also registered accordingly such that the email addresses were valid.

Phone Number

Same as email addresses, all resumes also had a phone number listed. To record the treatment effect, we assigned a unique phone number to each name we created. The phone numbers were purchased from Twilio, and were set up such that all calls and texts were redirected to a single Google Voice number. We first purchased the phone numbers manually and another scraping script was later used to make the purchases. The script randomly selected one from the states' area codes to filter the phone numbers before making the purchases to avoid too many numbers sharing the same area code. The purchased phone numbers were also manually checked and randomly tested by a research assistant. After the phone numbers were purchased and assigned to the names, they were listed in one randomly selected format of the two formats: "XXX XXX XXXX" and "(XXX) XXX-XXXX".

Volunteer Experience

We included one volunteer experience for every fictitious applicants. The volunteer locations were "Senior Center Kitchen", "Community Food Bank", "Soup Kitchen", and "Community Meal Program", and the responsibilities only included preparing and distributing food and cleaning for all locations. For each job we applied to, each fictitious resume used one of the four volunteer locations to avoid suspicion. The descriptions of volunteer responsibilities were paraphrased by ChatGPT and manually reviewed.

Template

For each job we applied to, the four fictitious applicants used four different resume templates. These templates were written in LaTeX.

B.1.3 Filing Online Applications

Birthday and Age

Each fictitious applicant had a high school graduation age between 17 and 19. If this applicant had no prior employment, we assumed the applicant's age last year was the same as the graduation age; if the applicant had prior employment, we assumed the applicant's age last year was the graduation year plus the length of work in years. To determine the applicant's birthday, we first determined their birth year by subtracting the year when the experiment was run by their current age (last-year age plus 1). Then, we added a randomly selected amount of days from uniform $[0, 364]$ to the first day of their birth year to arrive at their birth days. This method ensured all applicants were over 18 when applying, such that age limits would not confound the results.

SSN

In some cases, SSNs were required in the online applications. We assigned each fictitious applicant a SSN from the publicly available database of SSNs belonging to people deceased before 2007 (on National Archive (2007)). We manually selected different SSN files based on surname initials and varied starting digits of the SSNs to diversify the pool of SSNs assigned to the applicants. We also ruled out SSNs starting with 0. The steps requiring SSNs in the online applications are commonly hosted by third-party companies for verification purposes.

Questions

All online applications included some questions requiring responses from applicants. We only answered questions that were required and ignored all questions that were optional. To ensure the same answers across applicants, we asked the research assistants to fill in answers as the following: all daytime during the week (including weekends) for available date; a random day in the next two weeks from the application day for the start day; the average automotive sales salary ($35,000) for expected salary; having drivers licenses issued by their residential states (took "state-assigned" courses); not in any government subsidy programs such as SNAP; not wishing to disclose demographic information about gender, ethnicity, and veteran status; committing to only full-time roles; not having any certificate other than GED (if not high school diploma); having no felony history. Some application portals required at least one prior experience, and we used volunteer experience for candidates without prior experiences in such cases.

Application Details

To avoid suspicion and ensure the job listings are still open, we sent the four applications for each job during four separate morning/afternoon/evenings within three days. For example, suppose we sent the first application on Monday morning, the remaining three applications will be sent in arbitrarily selected three blocks in the following eight blocks of time: Monday afternoon, Monday evening, Tuesday morning, Tuesday afternoon, Tuesday evening, Wednesday morning, Wednesday afternoon, and Wednesday evening.

When applying, the research assistants also further checked for job requirements, whether the residential, prior employment (if any), and high school addresses matched the job location, and whether the prior employment (if any) was the same company as the job listing. Job listings with disqualifying requirements were discarded, and applications to jobs with disqualifying resumes were paused until all four resumes were ready to be sent. In the event of a job listing closing before we sent out all four applications, we use another job for the unsent conditions, such that the all four treatment arms were applied across the two jobs.

B.1.4 Local Market Thickness

For a group of companies, the general method for identifying automotive jobs within a municipality involved the following steps:

1. Filter all available job listings to include only full-time positions within a predetermined

radius around the city (5 miles). If a company website did not allow search within a five-mile radius, we manually looked up listings in towns within a 5 mile radius.

2. Scrape job listing URLs and titles from each page of the filtered results. To avoid redundancy, only one URL was randomly selected for each unique job title.

3. Visit each selected URL to extract the text outlining the job requirements. Keywords like ”required,” ”High School,” ”Ability,” ”level,” ”experience,” ”diploma,” or ”degree” were used to locate relevant information.

4. Assess job listings based on their education and experience requirements. Listings for managerial positions or those that required education beyond high school were disqualified.

5. Maintain a list of qualified job titles. If a newly scraped job listing had a title already present in the qualified list, it was automatically counted as qualified without further checks.

This general method required adjustments for certain companies due to variations in website structure or specific requirements. For instance, some companies' job listings might not have had a dedicated “requirements” section, requiring the use of different keywords and criteria to assess qualification. Other companies might have included remote work options or required bilingual skills, which needed to be excluded for the purpose of this analysis. The specific keywords and criteria used to identify qualified entry-level jobs were therefore tailored to each company to account for these variations.

Research assistants monitored the scraping process and randomly verified job titles. They also manually reviewed titles that were potentially not entry-level, such as ”manager,” and created lists of non-entry-level titles to exclude.

B.1.5 Call-Back Converstion Rate

To measure conversion rate from callbacks to employment, we recruited and surveyed 107 actual hiring managers from an online hiring platform from similar automotive industry stores as the ones targeted in our study. We first asked about their job titles and their duration in these titles. Then we proceeded to ask them about the conversion rate (between interview to actual job offer) in three blocks.

We first asked them a “base conversion” rate by providing no (experience or racial) information about the candidate: “Suppose you have reviewed an applicant's resume for a job that requires minimal experience, such as a cashier or inventory clerk. You have already decided to invite him or her for an interview. In your experience, what are the chances that the applicant ends up being offered the job? Note that 50% chance means that the person is offered the job half of the time, and 100% chance means the applicant is offered the job every time."

We then showed them the experience treatment block, where they were asked the conversion rates for two candidates: "Consider an applicant [with or without experience] who is interviewed. What is the chance that the applicant is offered a job after the interview?"

Lastly, we showed them the race treatment block, where we similarly asked about the conversion rates for another two candidates ("Consider a [Black or White] applicant who is interviewed. What is the chance that the applicant is offered a job after the interview?").

The order of the two candidates were randomized within each block, and in the race treatment Block, no experience information was mentioned for either candidate. Afterwards, the surveyed managers moved on to the demographic questions and one open-text question for them to leave comments.

B.1.6 Calculating Total, Systemic, and Direct Discrimination

B.2 Experimental IA

B.2.1 Preparing Letter Content

Focusing on STEM majors, we manually chose "Mechanical Engineering" and "Computer and Information Science" as the majors of the candidates. To avoid potential confounding effects of school rank and private universities, we chose large public universities similar in major rankings for each candidate: Penn State University for Mechanical Engineering and Ohio State University for Computer and Information Science. Each candidate had three prior experiences in their major-related field, where the job titles and descriptions were generated by ChatGPT and manually reviewed. We also used the InfoGroup 2022 business data to find their prior employers. We first filtered the data of all business based on location and primary SIC code (indicating industry) to only preserve companies whose industries are related to these two majors, and are located in nearby cities to the universities (Columbus, OH and Philadelphia, PA). We then used the first three companies for each city as the prior employers. Since the InfoGroup dataset does not seem to be particularly sorted based on any variables, being the first three companies was not suggestive of any attributes.

The durations of employment were determined randomly. For the most recent employment, we subtracted from the day of resume creation a random integer from uniform $[7, 28]$ as the most recent employment end date. This suggests all candidates had been unemployed for at most four weeks at the time of resume creation. For the second most recent employment end date, we further subtracted from the most recent end date a randomly selected number from uniform $[0.8, 1.2]$ times 365. This indicates that the time between the second and the most recent end dates were between 80% of a year and 120% of a year. For the earliest employment, we repeated the same process but subtracted from the second most recent end date. For the two most recent employment start dates, we added a randomly selected integer from uniform $[14, 45]$ to the previous end dates, indicating that the gaps between employment was between 14 and 45 days. For the earliest employment, we determined the start date by adding a random integer from uniform $[1, 60]$ to the same day of the month as the employment end date, but one year earlier and in July. For example, if the employment ended on Jan 15 2024, we added the random integer to July 15 2023. This is because the earliest employment should be the candidates' first employment after university, and our method would indicate this employment happened in the summer or fall of the graduation year. We then assumed all candidates graduated in early June of the same year as their earliest employment. The university start date was determined by first subtracting the university end dates by four years, and then adding to a random integer from uniform $[60, 100]$ such that university started in August or September.

For names of the candidates, we chose two male first names and two female first names. The four first names were mapped to four distinct common surnames. The first names are not suggestive of any minority race, and surnames all commonly belong to White people. The surnames were randomly selected from Kline et al. (2024) (*Bauer, Mast, Hostetler, Hershberger*), and the first names were chosen from the top four common names by gender during the 2000s listed by SSA to match the age of the candidates (*Emily, Olivia, Joshua, Matthew*). We grouped the names into two pairs, where each pair consisted of one male and one female name. Within each pair, the candidates shared the identical education and job experiences. The two pairs took up the two majors and corresponding university and experiences respectively.

B.2.2 Generating Example Letters

With the information, we asked ChatGPT to write recommendation letters by providing a prompt asking for a 300-word letter using all the education and prior experience information for each candidate. All such information was provided as bullet-in points similar to a resume layout. After generating the set of four letters, we checked for both word sentiment and word categories according to the LIWC dictionary (Kaplan et al. 2024). We verified that the letters displayed similar levels of bias in lexical content on the dimensions examined in Wan et al. (2023): letters written for male candidates were significantly more formal, positive, and agentic than those written for female candidates.

B.2.3 Generating multiple letters

Once we had the example letters, we changed the ChatGPT prompt to ask for similar letters in terms of sentiment and type of words for the same candidates and the same positions, which are outlined in the previous subsection. In this manner, we created 25 letters for each candidate. These correspond to the Endogenous-m and Endogenous-w letters for the male and female candidates, respectively. Next, we took 50 letters from the Endogenous-m set and replaced the names with their female counterparts', and the pronouns with she/her pronouns. These corresponded to the Exogenous-w letters.